393

14

STATISTICS: THEORY AND PRACTICE

BY

M. K. GHOSH,
M.A., B.COM. (Lond.), M. INST. T.
UNIVERSITY PROFESSOR AND HEAD OF THE DEPARTMENT OF COMMERCE, UNIVERSITY OF ALLAHABAD.

AND

S. C. CHAUDHRI, *M.A., B.COM.*
DEPUTY ECONOMIC & STATISTICAL ADVISER,
MINISTRY OF FOOD & AGRICULTURE, GOVERNMENT OF INDIA.
Allahabad Jubilee Gold Medalist.
Formerly, Lecturer, Allahabad University.

SEVENTH EDITION
(Thoroughly Revised and Enlarged)

ALLAHABAD
THE INDIAN PRESS (PUBLICATIONS) LTD.
1955

Price Rs. 12/8

Published by:
The Indian Press (Publications) Limited,
Allahabad.

Printed by J. N. Bose, at the Indian Press Ltd.,
93-A, Dharamtalla Street, Calcutta.

PREFACE TO THE SEVENTH EDITION

This book was first printed in November 1943. The second edition, which was revised and enlarged, was brought out in May 1945. The subsequent three editions, which were published in 1947, 1948 and 1949, were reprints of the second edition. The need for further revising and enlarging the second edition had long been felt, and the revised and enlarged sixth edition was brought out in 1954. The sixth edition was very quickly sold out and hence the need for the seventh edition of the book this year.

The changes made in the sixth edition were too numerous to be detailed. Three new chapters were added—on Regression, Business Forecasting and Probability and Sampling. Two old chapters were entirely re-written and brought up-to-date on Indian Statistics and Indian Index Numbers. Fresh topics were included in most other chapters, the more important of them being: Gross and Net Reproduction Rates, Quadratic and Progressive Averages, Splicing and Deflating of Index Numbers, Problems of Perfect Formula for Index Number Construction, Mixed Graphs, Zone Curve, Bank Curve, Zee Chart, Normal Distribution, Economics Curves, Theory of the Method of Moving Averages for Time-Series Analysis, Line of Best Fit, Parabolic Trend, Measurement of Seasonal and Irregular Fluctuations, Alternative Methods for Calculating Standard Deviation and Coefficient of Correlation, Coefficient of Correlation for Series of Cyclical Fluctuations, Coefficient of Correlation by the Method of Rank Differences, Scatter Diagram, Correlation Table, Consistence, the Chi-Square Test, Contingency and Historical Account of the Development of Statistical System in India. Besides, the maStter was rearranged in several chapters and the variety of illustrations and examples were increased to render the technical portions capable of being more

thoroughly grasped. Many of the old exercises were substituted by new ones. While making all these changes, every care was taken to retain the same design and purpose in the sixth edition as were set out for its predecessors.

We feel it necessary to add a word about the statistical data and index numbers relating to India. The quality and content of these statistics are gradually improving and the organization for collecting, processing and publishing them is undergoing changes. Our country is, thus, in a progressive stage so far as statistics are concerned. While we have taken every care to give the latest available inofrmation in respect of Indian statistics, including index numbers, in this edition, it would not be surprising if some of it goes out-of-date or becomes less important in the near future. The reader is, therefore, advised to keep himself in touch with the various sources of statistical data that we have indicated in the relevant chapters, so that he would get to know the changes in Indian statistics as they are made from time to time.

We crave the indulgence of the scrupulous reader for any error that may have crept in and shall be thankful for any suggestion to increase the usefulness of the book.

The latter of us holds, at present, the post of Deputy Economic and Statistical Adviser in the Directorate of Economics and Statistics, Ministry of Food and Agriculture, Government of India, but the book has no connection with his official duties and the Government are in no way responsible for any views expressed therein.

SUSHEEL CHANDRA CHAUDHRI

September, 1955. MOHIT KUMAR GHOSH

EXTRACTS FROM PREFACE TO THE SECOND EDITION

It has been an unexpected gratification to find that the response from the Universities and Colleges in India and other quarters to the first edition of this work was so hearty as to require a second edition of it in less than twelve months from publication. The fulfilment of this requirement has been delayed, among other reasons, by the scarcity of paper—so characteristic of the War-time conditions.

Opportunity has been taken of this new edition to make some additions, recast certain paragraphs and correct such errors as were noted in the first edition. The main additions are Indian Industrial Statistics, India's National Income, Method of Least-squares, Seasonal Variation Index, and Lag....

....

Reviews by journals and opinions from teachers of the subject and others interested in it have been very kind and, in some cases, equally suggestive. We are glad to avail ourselves of this opportunity of thanking them for their encouragement, and look forward to its continuance.

....

M. K. G.

S. C. C.

May, 1945

EXTRACTS FROM PREFACE TO THE FIRST EDITION

Statistics was once known as the Science of Kings, but now it has gained ground in almost every branch of human knowledge. For, the superstructure of human activity rests ultimately, if not primarily, upon a foundation of quantitative facts—facts, whose inherent complexity and confusion can be simplified and analyzed and which can be interpreted only with a knowledge of statistical methods. In this Age of Statistics, therefore, the importance of the study of the Science of Statistics cannot be over-emphasized, particularly for India whose development, in many spheres, is yet in its infancy......

This book is an attempt to furnish a simple, but comprehensive, text for those who desire to equip themselves with a knowledge of the elementary statistical methods to enable themselves to handle statistical problems like skilled workmen. ....

In a book such as this, the use of the viewpoints and materials of other works of parallel and higher standards is unavoidable. And, indeed, such works have been our valuable guide. But we have made every effort to so synthesize all these materials as to bring about unity and harmony. The treatment is non-mathematical, chiefly because the majority of those for whom this book is primarily meant are not expert mathematicians, and also because we feel there is a necessity of fundamental exposition of the non-mathematical, nonetheless vital, processes involved in statistical inquiries, analysis and interpretation......Once the readers have overcome their feeling of unfamiliarity and grasped the basic principles, it will be easy for them to pick up the higher and more mathematical statistics.......

Our thanks are due to Mr. Shiam Bahadur Kodasi, M.A., B. Com., who helped us in correcting the proofs.......

M. K. G.
S. C. C.

November, 1943

1827

CONTENTS

Chapter		Page
I	Background	1
II	Definition	13
III	Functions and Importance	22
IV	Planning a Statistical Inquiry	34
V	Collection of Data	42
VI	Editing the Collected Data	55
VII	Classification and Tabulation	64
VIII	Simple Derivatives	79
IX	Statistical Averages	90
X	Dispersion and Skewness	162
XI	Index Numbers	205
XII	Diagrammatic Representation	261
XIII	Graphical Presentation	304
XIV	Analysis of Time Series	368
XV	Correlation	403
XVI	Regression	460
XVII	Association of Attributes	480
XVIII	Interpolation	513
XIX	Business Forecasting	536
XX	Probability and Sampling	546
XXI	Interpretation of Data	598
XXII	Indian Statistics	610
XXIII	Indian Index Numbers	725
Appendix I A	Specimen of a Blank Form ..	791
Appendix I B	Specimen of a Questionnaire ..	794
Appendix II	Logarithm	797
	Mathematical Tables	I—XX

DETAILED CONTENTS

PAGES

CHAPTER I

BACKGROUND

Origin; Statistics and Mathematics; Statistics and Economics; Econometrics; Statistics in India. EXERCISES .. .. 1—12

CHAPTER II

DEFINITION

Definition of Statistics (data); Characteristics of Statistics; Statistical Methods; Science of Statistics; Functions of a Statistician; Main Divisions of Statistics. EXERCISES .. .. 13—21

CHAPTER III

FUNCTIONS AND IMPORTANCE

Functions of Statistics; Importance of Statistics; Scope of Statistical Science; Limitations of Statistics; Distrust of Statistics. EXERCISES .. 22—33

CHAPTER IV

PLANNING A STATISTICAL INQUIRY

Object and Scope of Inquiry; Nature and Type of Inquiries; Statistical Units; Degree of Accuracy. EXERCISES .. .. 34—41

CHAPTER V

COLLECTION OF DATA

Primary and Secondary Data; Primary Method; Choice of Enumerators; Training of Enumerators; Choice of Questions; Selection of Representative Data; Ensuring Collection of Reliable Data; Secondary Method; Precautions in using Secondary Data. EXERCISES .. .. .. 42—54

CHAPTER VI

EDITING THE COLLECTED DATA

Editing Primary Data; Accuracy; Approximation; Statistical Errors; Absolute and Relative Errors; Biassed and Unbiassed Errors; Possible Error; Editing Secondary Data. EXERCISES .. 55—63

CHAPTER VII

CLASSIFICATION AND TABULATION

CLASSIFICATION—Classification according to Attributes; Classification according to Class-Intervals; Statistical Series; Time, Spatial and Condition Series; Continuous and Discrete Series.

TABULATION—Rules and Precautions for Tabulation; Different types of Tabulation. EXERCISES 64—78

CHAPTER VIII

SIMPLE DERIVATIVES

Derivative defined; Subordinate Derivatives; Co-ordinate Derivatives; Purpose of Computing Statistical Derivatives; Derivative Series; Precautions in Computing and Using Derivatives; Mortality Rate; Fertility Rate; Gross Reproduction Rate; Net Reproduction Rate; Use of Simple Derivatives. EXERCISES .. .. 79—89

CHAPTER IX

STATISTICAL AVERAGES

Statistical Average and its Object; Law of Averages; Homogeneity of Data; Kinds of Average.

MODE—Location of Mode in Discrete Series; in Continuous Series; Advantages of Mode; Disadvantages of Mode; Uses of Mode; Graphical Measurement of Mode.

MEDIAN—Determination of Median in Individual Measurements; in Discrete Series; in Continuous Series; Advantages of the Median; Disadvantages of the Median; Uses of the Median.

QUARTILES, DECILES AND PERCENTILES—Location of Quartiles, Deciles and Percentiles in Individual Observations; in Discrete Series; in Frequency Distributions; Characteristics and Uses of Quartiles, etc.; Graphical Measurement of Median, Quartiles, etc.

ARITHMETIC AVERAGE—Simple Average in Individual Observations; in Discrete Series; in Continuous Series; Advantages of Simple Arithmetic Average; Disadvantages of Simple Arithmetic Average; Uses of Simple Arithmetic Average.

WEIGHTED ARITHMETIC AVERAGE—Computation of Weighted Average by Direct Method; What is Weighting? Computation of Weighted Average by Shortcut Method; When should Weighted Average be Used? Caution in Weighting; Difference between Simple and Weighted Averages; Characteristics and Uses of Weighted Average.

GEOMETRIC MEAN—Determination; WEIGHTED GEOMETRIC MEAN; Advantages of Geometric Mean; Disadvantages of Geometric Mean; Uses of the Geometric Mean.

HARMONIC MEAN—Determination; WEIGHTED HARMONIC MEAN; Characteristics and Uses of Harmonic Mean.

SOME IMPORTANT CONSIDERATIONS—Averages of the First Order; Typical and Descriptive Averages; Relative Position of Averages; Choice and Properties of Average; Limitations of Average.

QUADRATIC MEAN.
MOVING AVERAGE.
PROGRESSIVE AVERAGE.
STANDARDIZED DEATH-RATE.
EXERCISES .. .. 90—161

CHAPTER X

DISPERSION AND SKEWNESS

DISPERSION—Averages of the Second Order; Meaning of Dispersion; Measures of Dispersion.
METHOD OF LIMITS—The Range.

METHOD OF AVERAGING DEVIATIONS—**Mean Deviation** and its Co-efficient; Calculation of Mean Deviation and its Co-efficient; Characteristics and Uses of Mean Deviation and its Co-efficient. **Standard Deviation** and its Co-efficient; Calculation of Standard Deviation and its Co-efficient by Direct and Shortcut Methods; Characteristics and Uses of Standard Deviation and its Co-efficient; Modulus; Variance; Co-efficient of Variation. **Quartile Deviation** and its Co-efficient; Calculation of Quartile Deviation and its Co-efficient; Characteristics and Uses. Choice of Measures of Dispersion; Absolute and Relative Measures of Dispersion; Relation between Measures of Dispersion; Lorenz Curve; Practical Utility of Measures of Dispersion.

SKEWNESS—Tests of Skewness; Measures of Skewness; First Measure and Co-efficient of Skewness; Second Measure and Co-efficient of Skewness; Positive and Negative Skewness; Dispersion and Skewness Contrasted. EXERCISES .. 162—204

CHAPTER XI

INDEX NUMBERS

Definition; General Characteristics.
CONSTRUCTION OF WHOLESALE PRICE INDEX NUMBERS:

SELECTION OF ITEMS—Nature of Commodities; Number of Commodities; Quality of Commodities; Grouping of Commodities; Price Reporting Centres; Price Reporting Agency; Price Quotations; Frequency of Quotations; Averaging the Quotations.

CHOICE OF BASE AND CALCULATION OF RELATIVES—Fixed Base Method; Chain Base Method.
TYPE OF AVERAGE TO BE USED—Arithmetic Mean of Relatives in Fixed Base Method; Median of Relatives in Fixed Base Method; Geometric Mean of Relatives in Fixed Base Method; Chain Relatives; Merits and Demerits of Chain Base; Reversibility Test and Choice of Average.
SYSTEM OF WEIGHTING—The 'Unweighted' Index Number; Implicit and Explicit Weighting; Methods of Weighting; Summary; Limitations of Index Numbers.
COST OF LIVING INDEX NUMBERS: Difficulties in Construction; Construction; Aggregate Expenditure or Aggregative Method; Family Budget or Weighted Relatives Method; Group Indices; Errors in Cost of Living Index Numbers; Unsatisfactory Character of Cost of Living Indices.
INDICES OF INDUSTRIAL PRODUCTION.
BASE SHIFTING, SPLICING AND DEFLATING—Base Shifting; Splicing two Index Numbers; Deflating.
FISHER'S IDEAL FORMULA.
PROBLEM OF PERFECT FORMULA FOR INDEX NUMBER CONSTRUCTION—Developments in Index Number Theory; Important Formulae; Most Frequent Formula; Place of Chain Base; Uses of Index Numbers. EXERCISES. .. 205—260

CHAPTER XII

DIAGRAMMATIC REPRESENTATION.

Usefulness of Diagrams; Directions for Drawing Diagrams; Different Forms of Diagram; One Dimensional Diagrams—Simple Bar; Sub-divided Bar. Two Dimensional Diagrams—Rectangles; Squares; Circular Diagrams—Circles; Angular Diagrams—Sectors; Three Dimensional Diagrams—Cubes; Pictograms—Maps and Pictures; General Remarks. EXERCISES. .. .. 261—303

CHAPTER XIII

GRAPHICAL PRESENTATION.

Diagrammatic and Graphic Presentations Contrasted.
GRAPHS OF CONTINUOUS TIME SERIES—Rules for Drawing Graphs.
GRAPHS ON NATURAL SCALE—Absolute Historigram of one Variable; Index Historigram of one Variable; Absolute Historigrams of two or more Variables, (Homogeneous Units); Absolute Historigrams of two or more Variables, (Heterogeneous Units); Index Historigrams to compare changes of two or more Variables; Methods of Scale Conversion for comparing changes in two or more Variables; Mixed Graphs; False Base Line; Zone Curve; Band Curves; Zee Chart or Z Curve.
GRAPHS ON "RATIO" SCALE—Radio Scale; Logarithmic Curve; Instructions for reading of Logarithmic Curves; Advantages and Disadvantages of Ratio Scale; General Remarks.
FREQUENCY GRAPHS—Statistical Nature of a Group; Frequency Graphs for Discrete Series; for Continuous Series: Histogram; Frequency Polygon; Frequency Curve; Frequency Distributions and Shapes of Frequency Curves; Normal Distribution and Normal Curve of Error; Ogive Curve; Galton's Method of Locating the Median.
ECONOMIC CURVES—Law of Increasing Costs or Diminishing Returns; Law of Diminishing Costs or Increasing Returns; Demand and Utility Curves. EXERCISES. .. .. .. 304—367

CHAPTER XIV

ANALYSIS OF TIME SERIES.

Trend, Seasonal and Cyclical Fluctuations; Long-Period and Short-Period Changes; Measuring and Isolating Time Changes.
MEASUREMENT OF SECULAR TREND—Freehand Curve Method; Method of Moving Averages; Theory of Moving Averages Method; Limitations

of the Method of Moving Averages; Smooth Curve and Short-time Changes; Method of Least Squares; Mathematical Treatment of the Method of Least Squares; Characteristics and Limitations of Line of Best Fit; Parabolic Trend.
MEASUREMENT OF SHORT-TIME FLUCTUATIONS.
MEASUREMENT OF SEASONAL VARIATIONS—Seasonal Average; Seasonal Variation Index; Seasonal Variation.
MEASUREMENT OF IRREGULAR FLUCTUATIONS.
COMPARISON OF TIME CHANGES.
UTILITY OF TIME SERIES ANALYSIS. EXERCISES. .. .. 368—402

CHAPTER XV

CORRELATION

Meaning of Correlation; Positive and Negative Correlation; Linear and Curvilinear Correlation; Study and Measurement of Correlation.
COEFFICIENT OF CORRELATION—Degrees of Correlation; Limits of Co-efficient of Correlation; Karl Pearson's Formula; Pearson's Correlation Co-efficient by Direct Method; Alternative Direct Method; Pearson's Co-efficient by Short-Cut Method; Alternative Short-Cut Method; Co-efficient of Correlation for Long-Time Changes; Pearson's Modified Co-efficient for Short-Time Oscillations; Co-efficient for Cyclical Fluctuations; Correlation Co-efficient of Grouped Series; Alternative Method for S.D. of Grouped Series; Alternative Method for Co-efficient for Grouped Series; Assumptions of Pearsonian Correlation; Characteristics of Pearsonian Co-efficient; Probable Error of the Co-efficient; Interpretation of Co-efficient of Correlation; Correlation Co-efficient by Method of Rank Differences; Co-efficient of Concurrent Deviations.

CORRELATION BY GRAPHIC METHOD—Correlation Graph; Graphic Correlation of Time Changes; Lag and Lead; Scatter Diagram.
CORRELATION TABLE—Methods of Studying Correlation Compared. EXERCISES. .. .. 403—459

CHAPTER XVI

REGRESSION.

Ratio of Variation; Mathematical Computation of Ratio of Variation; Galton Graph; Reading the Galton Graph; Ratio of Regression; Usefulness of Galton Graph; Regression; Regression Line; Regression Equations; Co-efficient of Regression; General Remarks. EXERCISES. .. .. 460—479

CHAPTER XVII

ASSOCIATION OF ATTRIBUTES.

Statistics of Attributes; Notation and Terminology.
CONSISTENCE.
PROBABILITY AND EXPECTATION.
CRITERION OF INDEPENDENCE.
ASSOCIATION AND DISASSOCIATION—The Chi-Square Test; Independence, Association and Disassociation.
COEFFICIENT OF ASSOCIATION—Comparison of Intensity of Association.
PARTIAL ASSOCIATION.
CONTINGENCY—Manifold Classification; Contingency Table; Association in Contingency Tables, Co-efficient of Contingency. EXERCISES. .. 480—512

CHAPTER XVIII

INTERPOLATION

Meaning of Interpolation; Necessity of Interpolation; Assumptions; Accuracy of Interpolation; Methods of Interpolation.
GRAPHIC METHOD—Graphic Method in a Continuous Time Series; Graphic Method in Seasonal and Cyclical Series; Graphic Method and Correlation Curves.

ALGEBRAIC TREATMENT—Fitting a Parabolic Curve; Newton's Method of Advancing Differences; Lagrange's Formula; Forecasting; Conclusion. EXERCISES. .. .. .. 513—535

CHAPTER XIX

BUSINESS FORECASTING.

Business Conduct; What is Business Forecasting? Statistical Data and Methods used in Forecasting; Business Index or Business Barometer; Theories of Business Forecasting; Time Lag or Sequence Theory; Action and Re-action Theory; Cross-Cut-Analysis Theory; Specific Historical Analogy Theory; Underlying Assumptions; Utility of Business Forecasting; Limitations. EXERCISES. .. 536—545

CHAPTER XX

PROBABILITY AND SAMPLING

PROBABILITY—Elementary Principles of Probability; Simple Probabilities; Addition of Probabilities; Multiplication of Probabilities; Addition and Multiplication of Probabilities; The Binomial Expansion; Comparison of Actual and Theoretical Frequencies; Mean and Standard Deviation of Theoretical Frequencies; The Normal Curve of Error.

SAMPLING—Census *versus* Sample Inquiries; Objects of Sampling; Essentials of Sampling; Methods of Selecting a Sample; Bias; Avoidance of Bias; Sampling and Non-sampling Errors; Methods of Reducing the Random Sampling Error; Selecting a Random Sample; Tippett's Numbers; Conducting a Sample Inquiry; The Nature of Statistical Induction; Statistical Induction and Normal Distribution; Theory of Sampling.

SIMPLE SAMPLING OF ATTRIBUTES—LARGE SAMPLES—Mean and Standard Deviation in Simple Sampling of Attributes; Standard Error; Precision; Limitations of Simple Sampling.

SIMPLE SAMPLING OF VARIABLES—LARGE SAMPLES—Sampling Distribution; Simple Sampling of Variables; Standard Error of Arithmetic Mean; Test of Significance; Sampling Error; Median and Quartiles; Sampling Errors: Standard Deviation; Sampling Errors: Co-efficient of Correlation; Sampling Errors: Co-efficient of Regression; Sampling Errors: Differences between Means; Value of *T;* Limitations of Measures of Sampling Errors; General Remarks. EXERCISES. .. 546—597

CHAPTER XXI

INTERPRETATION OF DATA

Interpretation; Preliminaries to Interpretation; Mistakes due to False Generalisation; Wrong Interpretation of Index Numbers; Wrong Interpretation of Co-efficient of Correlation; Wrong Interpretation of Co-efficient of Association; General Directions for Interpretation. EXERCISES. .. .. 598—609

CHAPTER XXII

INDIAN STATISTICS.

Constitutional Position; No General Statistics Act; Recent Improvements; Statistical Organization in India.

STATISTICAL OFFICES UNDER THE GOVERNMENT OF INDIA—Ministry of Food and Agriculture; Ministry of Finance; Ministry of Commerce and Industry; Ministry of Labour; Ministry of Home Affairs; Ministry of Railways; Other Ministries; Central Statistical Organisation.

STATISTICAL ORGANISATIONS IN THE STATES.

POPULATION CENSUS—The Census of 1951; The 1951 and previous Censuses; Some Short-comings in Indian Censuses; Usefulness of Census.

VITAL STATISTICS.

AGRICULTURAL STATISTICS—Area Statistics—Temporarily Settled Areas; Area Statistics—Per-

manently Settled Areas; Yield Statistics—Traditional Method; Yield Statistics—Random Sampling Method; Land Utilization Statistics; Area and Production of Principal Crops; Crop Estimates; Shortcomings and Improvements in Agricultural Statistics; Season and Crop Reports; Livestock Statistics; Forest Statistics.

FOOD STATISTICS—Food Balance Sheet; Imports of Foodgrains.

INDUSTRIAL STATISTICS—Annual Census of Manufactures; Monthly Statistics of Industrial Output; Other Industrial Statistics; Collection of Statistics Act, 1953.

WAGES—AGRICULTURAL WAGES: Data published by Directorate of Economics and Statistics; Agricultural Labour Enquiry (ALE). INDUSTRIAL WAGES.

PRICES—Harvest Prices; Controlled Prices; State Bulletins; Bullion and Securities; Improvements Effected; Suggestions for Further Improvement.

TRADE STATISTICS—Monthly Foreign Trade Statistics; Foreign Trade by Sea and Air; Shipping; Foreign Trade by Land; Balance of Trade; Annual Foreign Trade Statistics; Other Foreign Trade and Allied Publications; Inland Trade Statistics; Raw Cotton Trade Statistics; Retail Trade; Review of Trade.

FINANCIAL STATISTICS—PUBLIC FINANCE: Central Budget; Revenue and Expenditure; Receipts and Disbursement. BANKING, CURRENCY AND EXCHANGE: Liabilities and Assets of Reserve Bank; Liabilities and Assets of Scheduled Banks; Liabilities and Assets of Non-Scheduled Banks; Other Banking Statistics; Co-operative Banking Statistics; Rural Credit Survey; Currency and Bullion.

NATIONAL INCOME—Problems of National Income Estimation in India; Methods of Measuring National Income; Method Adopted by the National Income Committee; National Income Estimates—

1948-49; Issues for the Future; Earlier Estimates of National Income.

NATIONAL SAMPLE SURVEY—The First Round of the Survey; Design of the Survey—First Round of N.S.S.; Poona Schedules.

GENERAL.

CRITICAL APPRAISAL OF INDIAN STATISTICS.
EXERCISES. .. .. 610—724

CHAPTER XXIII

INDIAN INDEX NUMBERS.

AGRICULTURAL PRODUCTION—Ministry of Food and Agriculture Index; Reserve Bank of India Index; The *Eastern Economist* Index; The F.A.O. Index; General Remarks.

INDUSTRIAL PRODUCTION—Interim Index of Industrial Production (Ministry of C. & I. Series); The *Eastern Economist* Index.

INDUSTRIAL ACTIVITY—*"Capital"* Index of Indian Industrial Activity.

COMMODITY PRICES—Index Numbers of Harvest Prices; Economic Adviser's Wholesale Prices Index; E. A's General Purpose *vs.* Sensitive Index; Index Numbers of Wholesale Prices in Calcutta; Index Numbers of Retail Prices (Urban Centres); Index Numbers of Retail Prices (Rural Centres).

SECURITY PRICES—Reserve Bank Index (Old Series); Reserve Bank Index (Revised Series).

WORKING CLASS COST OF LIVING INDEX NUMBERS—Labour Bureau Index Numbers for 15 Centres; Labour Bureau Index Number for Plantation Workers; States' Index Numbers; Bombay Working-Class Cost of Living Index; Kanpur Working-Class Cost of Living Index; Labour Bureau all-India Cost of Living Index; General Remarks.

EARNINGS OF FACTORY WORKERS—Labour Bureau Index of Earnings of Factory Workers in India; All-India Earnings and Cost of Living Indices Compared; General Remarks.

INDUSTRIAL PROFITS—Index Numbers of Industrial Profits.

FOREIGN TRADE—D.C.I. & S. Index of Unit Value and Quantum; Reserve Bank's Index of Unit Value & Quantum; D.C.I. & S. and R.B.I. Series Compared.

CONCLUSION—General Appraisal; Common Base Period for Official Index Numbers; Lacunae. EXERCISES. .. .. 725—790

Appendix I A	Specimen of a Blank Form. ..	791—793
Appendix I B	Specimen of Questionnaire. ..	794—796
Appendix II	Logarithm. ..	797—800
	Mathematical Tables ..	I—XX

STATISTICS: THEORY AND PRACTICE

CHAPTER I

BACKGROUND

The origin of the Science of Statistics can be traced to the days of yore when the ruling chiefs felt the need of having quantitative information about their man-power and material strength with a view to framing their military and fiscal policies and, to that end, ordered enumeration to be made of the population and property within their domain. Statistics was thus born as the Science of Kings and since its *raison d'etre* was the fulfilment of the needs of state administration, it came to be regarded as the Science of Statecraft. In fact, the word 'Statistics' itself seems to have been derived from the Latin *Status*, or, the Italian *Statista*, both meaning a political state. In the writings of Shakespeare and Milton the word *Statist* is found used in the sense of a person well-versed in state matters and helping high state officers responsible for framing government policies.

As time went by, different countries made progress in the political, economic and social field. With every such progress, the need for replacing quantitative expressions by quantitative statements was felt all the more keenly, since quantities made for better understanding of the given phenomenon. Accordingly, the scope of collection of facts and figures gradually got extended to several new fields. Thus, while during the Middle Ages the most common compilations in the West were concerned with taxation, distribution of land and available soldiers, much more comprehensive statistics relating to the social, economic and political sectors came to be collected during the Mercantilistic period when the policies of the western European governments were directed to the dual purpose of encouraging

such industries as enhanced the power of the state and of securing a favourable balance of trade.

In the 16th century, statistics found its use in astronomy. The astronomers compiled records of the motions of heavenly bodies and predicted about eclipses and positions of stars. Upon a study of the data collected by Tycho Brahe (1546-1601), Johannes Kepler discovered the three laws relating to the motion of planets on which the theory of gravitation was founded by Sir Isaac Newton.

The 17th century saw the study of vital and social statistics. In 1621, Professor George Obrecht of Strasburg University illustrated how vital and criminal statistics could be utilized for devising plans to provide a system of life insurance and pensions and to reform the criminals. Captain John Graunt of London (1620-1674) made an analytical study in the realm of vital statistics in 1661. Casper Neumann studied the death records of Breslau in 1691 and prepared his notes and conclusions, which fell into the hands of Edmund Halley, the famous astronomer and scientist, through the Royal Society of London. Halley computed from them a complete life table, deduced the expectation of life at each age and paved the way for a scientific system of life insurance. Sir William Petty (1623-1687) also drew up and discussed mortality tables. In 1698, the first life insurance institution was founded in London.

Statistics and Mathematics.

With many new fields opened up for statistical investigation, it was soon realized that the methods of collection and analysis of data, then prevalent, should be improved. The labours of Petty and Halley in the 18th Century had prepared the ground for a more scientific treatment of the statistical methods which they received at the hands of J. P. Sussmilch (1707-1767), a Prussian clergyman, who attempted to demonstrate the doctrine of 'natural order' statistically. Others devised statistical tables and graphs for effecting comparison of data. John Graunt, Petty and Sussmilch conducted their studies during the 17th and 18th centuries under the name of **political arithmetic.**

In the 18th century, an alliance was effected between Statistics and Mathematics and foundations were laid of the

Theory of Probability when J. Bernoulli (1654-1705), a professor of Basel, mathematically elucidated the 'law of large numbers' in his work *Ars Conjectandi,* published posthumously, and Daniel Bernoulli (1700-1782) suggested the theory of 'moral expectation'. The subject of probability grew out of an analysis of hazards of those who played games of chance. Laplace, the noted scientist, who followed up Sussmilch's statistical studies, raised the superstructure of the theory of probability in his creditable work published in 1812. And so did Gauss. But, it was left to the famous Belgian astronomer and mathematician, L.A.J. Quetlet (1796-1874), to lay the foundations of the modern theory of Statistics. Quetlet's meteorological researches brought him first to a study of vegetation, later to that of animal kingdom, and lastly to that of mankind. On a study of the physical, social and moral characteristics of men he found that every phenomenon yielded similar results; in each case there existed an 'average man' representing the average physical and mental qualities of society, and all other men, in respect of any particular character, diverged from the 'average man' with mathematical regularity. He found it to be true of all human actions, since he demonstrated that crimes, suicides, accidents all showed comparatively constant figures. He concluded that deviations from the 'average man' were subject to 'binomial' law and that the methods of probability, which had proved useful in treating errors of observation, could be profitably employed in Statistics. In fact, Quetlet recognized the significance of the principle of 'constancy of great numbers', which forms the basis of the modern theory of Statistics.

Several eminent mathematicians made their contribution to the Theory of Probability during the 19th century. Knapp (1842-1926) and Lexis (1837-1914) in Germany followed up Quetlet's principles and attempted a comprehensive study of the statistics of mortality. Sir Francis Galton (1822-1911), founder of the School of Eugenics, deserves the pioneer's honour among workers on Statistics for having conducted an enquiry into the principles relating to the transmission of mental and physical characteristics from one generation to another. His enquiry helped his great successor, Karl Pearson (1857-1936), to produce his notable work on biometry and to

emphasize the indispensability of statistics for the evolutionist as, in his opinion, the whole problem of evolution was a problem of statistics.

In fact, statistical methods constitute a branch of mathematics inasmuch as it is with the help of mathematical formulæ that the data are analysed, compared and interpreted. Averages, measurement of deviations from the norm, coefficients, curve-fitting, construction of index numbers are all essential to a comprehension of the collected data and they all have their foundation in mathematics. It is again the mathematical Theory of Probability which furnishes the basis for the 'law of large numbers' or the 'law of statistical regularity' on which is raised, as already said above, the entire superstructure of the modern theory of Statistics.

Statistics and Economics.

Statistics appeared rather late in the field of the Science of Economics, though a beginning was made by Sir William Petty in his work, *Political Arithmetic,* published in 1690, as also by Gregory King, who, about the same time, attempted to demonstrate statistically the relationship between supply of commodities and their prices. By the 18th century valuable statistical material relating to population, occupations, taxes, agriculture, industry, trade, shipping, etc., had been collected in most civilized countries; but there was no liasion whatsoever between statistical information and economic theory. Political economy was brought up in the Classical School, founded by Adam Smith, through his great work *Wealth of Nations,* published in 1776. Classical economists were staunch believers in the deductive and abstract method of reasoning, their lip-sympathy such as that held by J. S. Mill (1806-1873) to the advantage of statistical verification of deductive laws notwithstanding. W. S. Jevons (1835-1882), in his *Theory of Political Economy* published in 1871, also advocated verification of the deductive science of Economics by the inductive science of Statistics, and opined that political economy could be developed into an exact science if only commercial statistics were more complete and precise. Although Cournot (1801-1877), a renowned mathematical economist and writer on probability, did Statistics a signal service by hinting at the application of the calculus of

variations and making the first casual suggestion regarding the distinction between secular trend and periodic fluctuations, yet, it was W. S. Jevons who segregated seasonal movements, secular trends, and cycles, much as the modern writers do. Jevons' statistical work on prices was of a high order and he has been accorded the title of 'the father of index numbers'. He may be said to have put Statistics into Economics. In his *Theory of Political Economy* he wrote, 'I know not when we shall have a perfect system of statistics, but the want of it is the only insuperable obstacle in the way of making Economics an exact science.'[1]

The most emphatic weight on the introduction of Statistics into the study of Economics, however, came from the Historical School (1843-1883). This school, of which Roscher, Knies and Hildebrand were the representatives in Germany and Cliff Leslie in England, believed that economic doctrines were not to be reasoned out in the abstract but to be historically or inductively proved. The school, therefore, laid stress on history for past events and on statistics for the present ones. By the end of the 19th century, the attitude of a large body of economists towards the inductive method had become friendly. Alfred Marshall could write by 1907 that disputes as to the methods of study of Economics had ceased, that qualitative analysis had performed the larger part of its job, and that progress in the quantitative analysis depended upon the growth of realistic statistics. He asserted that induction and deduction were both needed for scientific thought, as the right and the left feet were both needed for walking. Pareto (1848-1923), whose work on *Political Economy* contained a comprehensive collection of statistics, expressed the conviction in 1907, that the progress of economic science depended largely upon the investigation of empirical laws derived from statistics. In F. Y. Edgeworth (1845-1926) could be found an economist, whose work on index numbers and correlation was particularly important and who greatly advanced the solution of statistical problems. Keynes, in his *Scope and Method of Political Economy* points out that the function of Statistics is "first, to suggest empirical laws, which may or may not be capable of subsequent deductive ex-

[1] 4th edition, page 12.

planation; and secondly, to supplement deductive reasoning by checking its results and submitting them to the test of experience."[2] It is now widely held that induction without deduction is barren and deduction without induction sterile.

Since the last quarter of the 19th century, two important factors have brought about a fundamental change in the place of Statistics in Economics. Firstly, pure theory of Statistics has made a remarkable improvement. Eminent men like August Meitzen, Francis Edgeworth, Francis Galton, Karl Pearson, G. Udny Yule, C. B. Davenport, W. S. Gossett, A. L. Bowley, Adams, W. Persons, W. I. King and R. A. Fisher have developed statistical methods like probability, sampling, curve-fitting, correlation, time series analysis, interpolation and index numbers. Secondly, a large mass of statistical data has been collected in different countries of the world and foundations have been laid for their regular and scientific collection in future too. This has been made possible by the establishment of statistical bureaus and scientific recording of varied economic statistics such as population, production and other censuses, prices, wages and trade staitstics. These economic statistics perform the dual function of verifying economic theory and furnishing data for the guidance of economic policy or practice. The theory of Economics attempts to explain the action of man in the pursuit of material well-being. It shows how under the influence of wants man proceeds to their satisfaction. The actual behaviour of man is shown by a study of the economic organization. Statistical observation is one method of studying the economic organisation, and therefore of confirming the adequacy of the theory of Economics as an explanation of human behaviour. Thus, economic statistics are brought into relation with economic theory. Further, the actual life of any community is constantly affected by administration, legislation, public opinion, etc. Every community has its economic policy which is directed to the solution of the economic problems facing that community. But policy can have no sure basis unless it is framed on an analysis of the facts of economic life and a study of the results of economic action. Statistics furnish one method of measuring and analysing the given phenomena and

---

[2] 2nd edition, page 338.

of tracing the effects of economic action. In this manner, statistics help in the framing of economic policy. Thus both in economic theory and practice, the recourse to the statistical method is increasing. Indeed, as Marshall said while giving evidence before the Committee of Census, "Statistics are the straw out of which, I, like every other economist, have to make the bricks." It may even be said that the Science of Economics is becoming statistical in its method.

Econometrics.

Economics, Mathematics and Statistics have in recent years got so much inter-related that a new science by the name of Econometrics has now been evolved with the object of harnessing the three sciences to the study of a given phenomenon. The modern science of statistics is thus no longer synonymous with 'political arithmetic' or 'science of statecraft'. Far beyond the matters of state, the scope of the modern science has been extended to varied departments of human knowledge, to all those phenomena where quantitative studies are involved.

Statistics in India.

In India, as elsewhere, statistics came to be collected as facts of interest to the statesman. In this sense they were compiled as long ago as 300 B.C. during the rule of Maurya Kings, and subsequently in the reigns of Asoka, Gupta Dynasty and Moghul rulers. Kautilya's *Artha-Shastra* is replete with statistics of land, prices, wages, population, etc. collected during Maurya rule and *Ain-i-Akbari* with those relating to Akbar's reign (1556-1605 A.D.). With the gradual assumption of political power in the latter half of the 18th century the East India Company also started collecting statistics, particularly those relating to agriculture. Their bulk increased with the adoption of the Ryotwari System in the country and with the transference of power from the Company to the British Crown after the Mutiny of 1857. A *Statistical Abstract relating to British India* was prepared for the first time in 1868. It continued to be published in England as an annual feature till 1923 when it was transferred to India. In 1881, the first edition of the *Imperial Gazetteer of India* was brought out which contained economic statistics of the different parts of the country. In 1881 a population census was conducted, which was the first

complete and synchronous enumeration in India. The census data were included in the second edition of the *Imperial Gazetteer*, published in 1885-87.

In the last quarter of the 19th century various departments of the Government of India started bringing out statistical publications relating to their subjects. Following the recommendations of the Famine Commission of 1880, Departments of Agriculture were created in the Provinces in 1881. In December 1883, a Statistical Conference was held in Calcutta, at which it was *inter alia* decided that the Provincial Governments should prepare Crop Forecasts and conduct Quinquennial Livestock Census. All-India Crop Forecasts started with rice and wheat in 1894 and by 1900 oilseeds, cotton, jute and sugarcane had been covered. Thereafter the number went on increasing steadily. The first Quinquennial Livestock Census was conducted in 1887-88. Statistics of agricultural wages began to be collected in 1873 on half-yearly basis and were published in the Provincial *Gazettes*. They were also published on annual basis in the *Prices and Wages*. A publication entitled *Returns of Agricultural Statistics of British India* was brought out for the first time in 1886 by the Revenue and Agriculture Department and continued as an annual feature. The statistics of foreign trade, on the other hand, were being published by the then Finance and Commerce Department. A Statistical Bureau to deal with the statistics of both these Departments was set up in 1895 headed by a Director-General of Statistics. In 1905, the Director-General of Statistics was replaced by a Director-General of Commercial Intelligence, who was to maintain liaison between the Government and the trading community besides continuing the work formerly done by the Director-General of Statistics. He brought out the first issue of the *Indian Trade Journal* in 1906. In 1912, the headquarters of the Government of India were shifted from Calcutta to Delhi and it was then decided to separate statistics from commercial intelligence. This separation was, however, found to be inconvenient and statistics and commercial intelligence were amalgamated again in 1922. The combined Department was named as the Office of the Director-General of Commercial Intelligence and Statistics and continued to be located in Calcutta.

In 1925, an Economic Enquiry Committee was set up under the chairmanship of Sir M. Viswesvarayya to enquire into "the question of the adequacy of the statistical data available and the desirability and possibility of supplementing it and of undertaking an economic enquiry". In their report the Committee *inter alia* observed that to maintain Indian statistics on a satisfactory basis the statistics of all the Central and Provincial Departments should be placed under the supervision of one Central Authority, and that each Province should have a Statistical Bureau. The Royal Commission on Agriculture, 1928, agreed with the latter part of the Committee's recommendation but assigned to the Central Department of Statistics the role of only co-ordinating the Provincial statistics with a view to bringing about uniformity in them. The Royal Commission on Labour, 1931, emphasised that it was on facts that policy must be built and recommended that recourse should be taken to legislation to enable the collection of information from employers regarding wages and living conditions of industrial labour, from merchants regarding prices, from land-lords regarding rent, etc.

A wholesome outcome of the recommendations of the Royal Commission on Agriculture was the setting up of a Statistical Section in the newly opened Imperial (now Indian) Council of Agricultural Research. The publication relating to statistics of internal trade was revived and collection of statistics of industrial production on a voluntary basis was started. In 1933, a Statistical Research Bureau was set up in Delhi to analyse and interpret economic statistics. In 1934, Messrs. Bowley and Robertson conducted an enquiry into the possibility of an Economic Census of India. They recommended *inter alia* the appointment of a permanent economic staff including a Director of Statistics at the Centre. This organization was to be made responsible for conducting Census of Production and Census of Population and the co-ordination of Central and Provincial statistics. In pursuance of these recommendations, the Government of India decided in 1936 to set up a central organisation at Delhi but this could not be achieved immediately owing to financial stringency. An Office of the Economic Adviser to the Government of India was, however, created in 1938 for the collection and study of information bearing on

India's economic welfare and advising the Government as a whole on economic matters. The Statistical Research Bureau was merged with this Office.

After the World War II broke out in 1939, the need for collecting comprehensive data in various fields was keenly felt. Both at the Centre and in the Provinces, several departments set up their own statistical organisations to collect, compile and publish statistical data and tender advice on their basis to the Government. Statistics relating to land utilization, crop production and plantations, however, continued to be collected and published by the Director-General of Commercial Intelligence and Statistics till a Directorate of Economics and Statistics was set up under the Ministry of Food and Agriculture at Delhi in the beginning of 1948 and took them over from the Director-General. The Directorate is also concerned with the collection, compilation and publication of other agricultural statistics, such as prices of agricultural commodities, agricultural wages, forest statistics, statistics relating to food distribution, Grow More Food campaign, sugar production, etc. Several other important developments have also taken place in recent years, particularly after Independence since when the growing emphasis on economic planning has been calling for greater and more reliable statistics. Emphasis has, for some time, also been laid on effecting improvements in the quality and content of existing statistics as well as collecting new data. Armed with the Industrial Statistics Act, 1942, the Directorate of Industrial Statistics conducted the first Census of Manufactures in 1946 and has been continuing it annually since. The Labour Bureau started the construction of Working-Class Cost of Living Index Numbers for certain urban areas on a uniform basis with 1939 as the base and Retail Price Index Numbers at certain rural centres with 1944 as the base year. The Economic Adviser's Sensitive Wholesale Price Index Number of 29 articles was substituted by the General Purpose Index Number with effect from January, 1947. A National Income Committee framed the first official estimate of the national income of India for 1948-49 in the year 1951. A Statistical Unit was set up in 1949 under the Cabinet Secretariat at Delhi, with Professor P. C. Mahalanobis as the Statistical Adviser, to co-ordinate all activities relating to statistics. This Unit developed into a

Central Statistical Organization at the beginning of 1950. Some of the States, like Madras, West Bengal, Bihar, U.P., Madhya Bharat, Rajasthan have centralised the collection of the larger bulk of State statistics under their newly created statistical organisations.

Along with the expansion of statistics and statistical organisations, attention in recent years has also been devoted to research and training in statistics. The Indian Statistical Institute at Calcutta, and the Indian Council of Agricultural Research at Delhi have, for instance, evolved the technique of random sample surveys for the estimation of yield of crops. These institutions also impart training in statistics. More recently, Indian nationals have been sent abroad also for acquiring training there. A session of the International Statistical Conferences was also held in India in December 1951, and was followed by a short training course. Statistical Quality Control in industry is a new development in the Science of Statistics, to which too thought has recently been devoted in the country.

Several improvements have also taken place in recent years in the methods of collecting data. The 1951 Census of Population, for instance, was a great improvement over the previous censuses. The technique of random sampling evolved by the Indian Council of Agricultural Research is now being largely employed by various States to estimate the yield of crops in substitution of the "annawari" method. West Bengal is utilizing the sampling technique developed by the Indian Statistical Institute for the same purpose. The all-India Agricultural Labour Enquiry conducted in 1949-50 and 1950-51 by the Ministry of Labour, the all-India Rural Credit Survey launched in 1951-52 by the Reserve Bank of India and the multi-purpose National Sample Survey being conducted since 1951 by the Ministry of Finance are all based on the random sampling technique. These developments go to show that now the statistics in India are not simply the 'by-products of administration' which the Bowley-Robertson Committee found them to be in 1934; they are steadily preparing themselves for serving the needs of a Welfare State.

As in most other countries, official statistics constitute by far the greatest bulk of statistical material available in India.

Non-official or semi-official institutions collecting and publishing statistical data are very few. They include bodies like the Boards of Economic Enquiry, Universities, Research Institutes, Chambers of Commerce, Manufacturers' Associations and Economic Journals. A few individuals have also conducted some economic enquiries. The non-official data, however, generally pertain to local investigations and hardly constitute any regular continuous series except in the case of some of those relating to business, trade or commerce. The use of statistics outside the Government is, however, rapidly increasing. In business houses, banks and insurance offices, resort is being increasingly taken to the utilisation of statistical devices for planning and measuring the progress. Non-official bodies like the Congress Agrarian Reforms Committee and members of the Central and State Legislatures have also been falling back on statistical data. A general interest in statistics has thus been aroused which is bound to lead to greater development of statistics in the country.

EXERCISES

(1) Trace the growth of the science of statistics, and throw light on its future.

(2) Explain the relationship between Economics and Statistics. How far has the use of statistical methods in Economics led to its development?

(M.A., Agra, 1952).

(3) 'Statistics are the straw out of which, I, like every other economist, have to make the bricks.' (*Marshall*).

Explain, in the light of the above observation, the relation between Economics and Statistics, and discuss how far it is correct to say that the Science of Economics is becoming statistical in its method.

(M. Com., Alld., 1944).

(4) Statistics was born in the needs of state administration, but is no longer concerned with matters of state alone—In the light of this statement, show how and what transformation has taken place in the meaning of statistics in India.

(5) There could be no Statistical Theory without Mathematics, yet that Theory is no more a branch of mathematics than is, say, engineering or astronomy.—Discuss.

CHAPTER II

DEFINITION

Statistics* are numerical statements of facts in any department of enquiry, placed in relation to each other[1]. All figures are not necessarily statistics. 20, 38, 67, or 15, 32, 55 are undoubtedly quantitative figures, but not statistics. For, neither do they concern a sphere of inquiry, nor are they placed in relation to each other. If, however, we are told that for husbands, in a certain community, at ages 20, 38, 67 years and so on the corresponding ages of wives are 15, 32, 55 years and so on, these figures at once become statistics. For, now they throw light on the relationship between the ages of husbands and wives in the given community.

Characteristics of Statistics.

Several facts emerge from the above. Firstly, **statistics must be quantitatively expressed.** Qualitative expressions like 'young', 'middle-aged' and 'old' have been indicated by numerical expressions like 20, 38, and 67 or 15, 32, and 55 years. Likewise, yields of crops over a series of years expressed in maunds per acre are statistics, but expressed by such terms as 'good', 'fair', 'normal', or 'poor', are not. Secondly, **statistics are always aggregates.** A single age of 20 or 33 years is not statistics; a series of ages are. A single birth, a sale or a consignment does not form statistics. Yet numbers of births, sales or consignments are statistics, since they can be studied in relation to time, place and frequency of occurrence. Thirdly, **statistics**

---

* The term statistics is applied to the science of statistics as well as to its subject-matter. In the former sense it is used as singular, and in the latter as plural, noun.

[1] Bowley, A. L., *An Elementary Manual of Statistics*, page 1.

should relate to a department of inquiry. That is, the sphere on which they are to throw light must be definite and clear. Their purpose and object must be predetermined. The purpose of a series of ages of husbands and wives in our example may be to find whether young husbands have young wives and the old, old. Fourthly, **statistics must be capable of being placed in relation to each other.** That is, they should be comparable. To be so they should be homogeneous; they cannot be stray numerical facts, unrelated data, culled from indiscriminate sources having no common basis for selection. Ages of husbands are to be compared only with the corresponding ages of wives in our example, and not with, say, heights of trees.

But these are not the only characteristics which statistics possess. It should be added that **statistics are affected to a marked extent by a multitude of causes.** They are hardly ever traceable to a single cause. They are related to other facts. The stature of a man, for example, is causally connected with a number of factors like his race, ancestry, diet, age, occupation, climate and habit. It may be further added that **reasonable standards of accuracy must be observed in the enumeration or estimation of statistics,** if logical conclusions are to be based on them. Of course, the reasonableness of the standards will be determined by the purpose which statistics are to serve. If precise results are desired, statistics must be accurately compiled; but if only general impressions are looked for, even rough estimates may serve the purpose.

A definition, comprehensive enough to include the above characteristics of statistics (in the sense of statistical data), is given by Secrist,[2] who uses the term statistics as meaning **"aggregates of facts, 'affected to a marked extent by a multiplicity of causes,' numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose, and placed in relation to each other."**

Statistical Methods.

These statistics constitute the raw material which must

---

[2] Secrist, H., *An Introduction to Statistical Methods*, 1933 ed., page 10.

pass through certain mechanical processes to yield finished products. The most important step to be taken is to reduce the multiple causes affecting the data to a comparatively small residuum. For this purpose, the experimental method has been carried to perfection in 'exact' sciences like physics and chemistry, where the data are capable of being measured with reasonable accuracy, the salient factors of the problem are few and simple, and conditions are within the experimentor's control. These circumstances favour the application of experimental method. Experiment has the merit of replacing complex and varied systems of causation by simple ones, where only one cause would vary at a time. But, the experimental method proves unsatisfactory and inadequate in 'inexact' sciences like biology and sociology, where the data are mixed with extraneous and irrelevant matters, factors of the problem are many and complex, and conditions are not under control. These circumstances do not permit experiment. The inexact sciences, denied simplification of data through the experimental method, have in general to deal with data as they occur—data affected to a large extent by a number of alternative causes acting jointly or severally. They apply some method other than the experimental one to render their complex and unwieldy data intelligible. To serve this end, important and more persistent factors influencing the data have to be segregated from the casual disturbances that cancel out in the long run, and the extent to which the observed effect results from the operation of each one of the former factors has to be studied. The collected figures are classified, tabulated, compared, correlated and finally interpreted. Methods employed in these different processes are termed 'Statistical Methods'. Thus, **statistical methods are the devices by the application of which quantitative data influenced by multiple causation are collected and scientifically analysed and elucidated so that they are brought within easy and clear grasp.**

It should not be understood that the machinery of statistical methods is employed only in inexact sciences. Statistical methods may also be profitably used in exact sciences, for whatever the perfection of the experimental devices they can hardly ever be *absolutely* perfect. The observer of physical or chemical phenomenon and the instrument of observation are both

sources of error; and the effects of changes in moisture, pressure, temperature, etc. cannot be totally eliminated. Statistical methods are, therefore, the handmaid of both the exact and the inexact sciences, but are of greater service to the latter. Experimental method is, of course, more precise than the statistical one, but the latter often affords fairly successful results when the former fails. A scientific exposition of the statistical methods is named as the Theory of Statistics.

Science of Statistics.

Different authors have given different definitions of the Science of Statistics, emphasizing different aspects. **Webster** defines Statistics as "Classified facts respecting the condition of the people in a state . . . especially those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement."[3] This definition, however, relates to the use of the term statistics in the sense of statistical data rather than in that of the science, and even so is much in keeping with the original meaning of the term. In its modern sense the term is not confined to 'the condition of the people in a state', but has stretched itself to almost every phenomenon—biological, astronomical, social, meteorological—where a study of large numbers is involved. Webster's definition is therefore inadequate.

According to **Bowley** "Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations."[4] This definition, according to its author, concerns the student of sociology, political economy or demography. But when the author recognizes that 'Statistics is not merely a branch of political economy nor is it confined to any one science',[5] its definition should not have been so drawn up as to limit its operations to only one field—*viz.*, that of man and his activities. This definition is, therefore, not sufficiently inclusive. Bowley further observes: 'Statistics may rightly be called the science of averages.'[6] No doubt, averages present a bird's-eye view of a mass of unintelligible data, but there are

---

[3] Quoted by King, W. I., *Elements of Statistical Methods*, page 20.
[4] Bowley, A. L., *Elements of Statistics*, page 7.
[5] *Ibid.*, page 4.
[6] *Ibid.*, page 7.

other equally serviceable devices, such as graphs, pictograms, correlation tables and co-efficients, which modern statistics utilizes to comprehend the significance of the complex quantitative data. Therefore, while this definition does not confine the scope of the science to a particular phenomenon, it is still inadequate in so far as it stresses only one of the several statistical methods. Suggesting a possible definition Bowley says that statistics may be called "the science of counting".[7] Analysing this definition he observes that while dealing with large numbers, such as a population census, counting is neither easy nor within the reach of an individual. Great numbers, instead of being counted, are estimated. But because of varying degrees of intelligence and sense of accuracy among the estimators, and also because of the difficulty of so clearly defining the object to be counted that every worker shall understand the same thing by the same definition, the estimates are not mathematically exact. Bowley then concludes that 'though all estimates of this nature are sometimes included under the term *Statistics*, this definition at once is too wide, and also does not bring out the distinctive nature of statistical method.'[8] Therefore, this definition is also far too restricted.

Boddington denominates statistics as a 'science of estimates and probabilities'.[9] This is vague and indicates a narrow point of view inasmuch as estimates and probabilities are only a part and not the whole of statistics.

King defines statistics thus: 'The science of statistics is the method of judging collective natural or social phenomena from the results obtained by the analysis of an enumeration or collection of estimates'.[10] The author himself regards it as possible that statistical problems such as would fall outside the limits of this definition might be imagined but maintains that it is sufficiently broad for practical purposes.

In order that the definition of the Science of Statistics may suit its modern sense, it should be so framed as to include all that is rightly its and to exclude what is extraneous to it.

[7] Bowley, A. L., *op. cit.*, page 3.
[8] Bowley, A. L., *op. cit.*, page 4.
[9] Boddington, A. L., *Statistics and their Application to Commerce*, page 7.
[10] King, W. I., *op. cit.*, page 23.

Enough has already been said of what statistics and statistical methods are, yet a few observations are necessary before arriving at a suitable definition. Statistics is concerned with mass phenomena, with large numbers descriptive of groups, with results of collective action. Individual facts and figures may be of interest to an individual: Statistics does not deal with them. For example, an individual's income is not statistics, nor does the gathering of knowledge about an individual's income become a statistical study. If, however, we compare the total earnings of a group of workers with other elements in a business, say with its turnover, we arrive at a clearly defined relationship between earnings and turnover. This relationship should hold good in normal circumstances. The study of this relationship is a statistical study. Individual peculiarities count for nothing in it. It is the possession of the same peculiarities by the whole or a majority of the constituents of the group that is significant. Further, for analysing the data influenced by a multitude of causes recourse has to be taken to methods that are based on particular characteristics of large numbers. For instance, a characteristic of large numbers and of averages derived from them is that they enjoy great inertia; individual incomes may change very fast, total or average incomes vary very little. Through such methods accuracy of statements is examined, complicated data are analysed and one estimate is compared with another. All those estimates to which these methods apply fall within the scope of Statistics irrespective of the fact whether the estimates concern the field of sociology, biology, physics or any other study. Statistics is, therefore, not confined to any particular branch of human knowledge; it is all-pervading. It may, therefore, be concluded that **the Science of Statistics is a study of the methods applied in collecting, analysing, and interpreting quantitative data, affected by multiple causation, in any department of inquiry.**

Functions of a Statistician.

The functions of a statistician are then simple. He is concerned, firstly, with the collection of statistical data, secondly, with their analysis and finally with the interpretation of the results of such analysis. Sometimes a sort of division of labour may be noticed in that a statistician may be engaged only on

the collection of data, leaving their analysis and interpretation to another. But such a division of duties may not always result in the best elucidation of the given problem.

A statistician cannot work miracles. He is not an alchemist expected to produce gold from any worthless material. He is rather like a chemist capable of assaying the value the material contains and of extracting nothing more than this value. It would, then, be no use commending a statistician because his results are precise nor condemning him because they are not. If he is gifted with the competence his craft demands, the value of his results shall follow solely from the material he analyses. His job is only to produce what the material contains, and no more. He must, therefore, be an impartial umpire, free from fear or favour. His personal prejudices should not affect the conduct of his daily duties. He should be gifted with sound commonsense and exercise caution in stating the conclusions emerging from the data.

Main Divisions of Statistics.

The domain of statistics can be generally classified into two main divisions: Statistical Methods and Applied Statistics.

Statistical Methods are concerned with the formulation of the general rules and principles applicable in handling different branches of data, *e.g.*, the methods of collection of data, classification, tabulation, comparison by means of averages, diagrams and coefficients, correlation, etc.

Applied Statistics deals with the application of these rules and formulæ to concrete subject-matter like wages, prices, trade, population. Applied Statistics may consist of biometry, psychometry, vital statistics, administrative, social and economic statistics.

EXERCISES

(1) Statistics are not a mere mass of figures.—Elucidate.
(M.A., East Punjab, 1952).

(2) 'The statistical method is in a sense a substitute for the method of controlled experiment—a poor substitute, but the best we have.'

Explain the above statement with examples and comments.
(M.A., Alld., 1950).

(3) 'By statistics we mean quantitative data affected to a marked extent by a multiplicity of causes' (Yule)—Explain.

(4) Statistics are "aggregates of facts 'affected to a marked extent by multiplicity of causes', numerically expressed, enumerated, or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose, and placed in relation to each other."

Elucidate the above definition, bringing out clearly the characteristics of statistics.

(B.Com., Alld., 1945).

(5) Comment on the following definitions of Statistics—

(*a*) By Theory of Statistics or, more briefly statistics we mean the exposition of statistical methods.

(*b*) The theory of statistics comprises an analysis and interpretation of systematic collection of numbers relating to the enumeration of great classes.

(*c*) Statistics is the science of estimates and probabilities.

(*d*) Statistics is the science of counting.

(6) Discuss the scope of the Science of Statistics and show its relation to natural and social sciences.

(B.Com., Lucknow, 1940).

(7) Write a short essay on the statistician's role in post-war reconstruction.

(M.A., Agra, 1944).

(8) 'Statistical methods include all those devices of analysis and synthesis by means of which statistics are scientifically collected and used to explain or describe phenomena either in their individual or related capacities.' (Secrist)—Elucidate the above statement.

(B.Com., Nagpur, 1945).

(9) By statistical methods we mean methods specially adapted to the elucidation of quantitative data affected by a multiplicity of causes. (Yule)—Comment.

(10) What are the important duties of a statistician? Under what conditions would he be successful in his mission?

(M.Com., Rajputana, 1951).

(11) (*a*) A statistician is not an alchemist expected to produce gold from any worthless material.

(*b*) 'Statistics is the Science of the measurement of the social organism, regarded as a whole, in all its manifestations'.

(*c*) Statistics are numerical statements of facts in any department of inquiry, placed in relation to each other.

Comment on any two of the above statements.

(M.A., East Punjab, 1951).

(12) Discuss the claims of Statistics to be regarded as a science.

(M.A., Agra, 1930).

(13) 'Statistics affects everybody and touches life at many points. It is both a science and an art.'

Explain the above statement, with appropriate examples.

(B.Com., Alld., 1952).

CHAPTER III

FUNCTIONS AND IMPORTANCE

Functions of Statistics.

Statistics simplifies complexity. Human mind is unable to assimilate a mass of complicated data at a stretch, but if they are simplified such that their fundamental features are brought to light they at once become understandable. One can hardly form an unquestionable opinion regarding the comparative examination standards of two universities if he were told the marks obtained by every one of the two thousand students of each university. But if these unwieldy data were reduced to number of passes, percentage success, class-wise or division-wise success, or presented through diagrams they would become readily intelligible. The Science of Statistics does this job. It boils down complex data to simple representative numbers easily adaptable to human mind.

Statistics enlarges individual experience. One may exercise his best ability and power of judgment to view the quantitative significance of a phenomenon; for instance, one may attempt an estimate of the rise in cost of living in a town at a particular time, but in the absence of adequate statistical data such estimate would be a mere guess, being subject to vagueness, inaccuracy and personal prejudices. If, however, one were to first conduct an investigation and collect the requisite data, the resulting estimate would not only correct the former conjecture, but would also, in the process, leave a lasting impression on the investigator's mind.

Statistics compares the simplified data and measures their relationship. To appreciate the meaning of one estimate we often need another for comparison. A statement of wholesale

prices of different commodities over time in a country may, in itself, not carry much meaning. It may, however, be compared with the volume of money in circulation over the same time to study the relationship between the course of prices and money supply. To successfully do so, both the prices and money-supply data should first be reduced to a comparable basis, say, to index numbers. The study as to how prices, on the whole, react to changes in money supply then becomes easy.

Statistics tests the laws of other Sciences. Deductive laws of several sciences can be verified with the help of relevant statistical data and application of statistical methods.

Statistics guides in the formulation of policies. Scientific analysis of statistical data constitutes the starting point in all policy-making. How much of foodgrains should India import in a given year or what quantities of jute goods can it spare for export abroad, are questions whose answers are to be found only on a study of related statistics. Statistics are the straw out of which an economist makes bricks.

Statistics enables realization of magnitudes. That India's population in 1951 was 36 crores or that her normal production of cereals is 44 million tons a year are statements which enable one to visualise the whole country so far as its population or cereal production is concerned. More than that, if one were told that the population in India increased on an average at the rate of 4.2 million persons annually during 1941-51, or, better still, at the rate of 11,500 persons per day or 8 persons per minute, an idea of the magnitudes involved will become more clear to him.

Importance of Statistics.

Statistics is the Arithmetic of Human Welfare. It was once termed as the Science of Kings. In ancient times it kept the kings informed about the man-power and riches of their domain. What are now called statistical studies were in the past conducted under the name of *Political Arithmetic*. Civilisation has much advanced since then, and the application of the mathematical theory of probability to social phenomena has yielded an ingenious apparatus to deal with the figures of wealth and welfare. Statistics is indispensable these days for a clearer appreciation of any problem affecting the welfare of

mankind. Problems relating to poverty, unemployment, food shortage, protective tariffs, uneconomic agricultural holdings, etc., cannot be fully weighed without the statistical balance.

Planning without statistics cannot be imagined. Neither can a policy be scientifically chalked out, nor can its success be measured without the statistical apparatus.

Statistics discloses casual connection between related facts. Such study is at the bottom of all sound human endeavour. Statistics is the light-bearer that enlightens the way to life's adventures. It unravels the crowded complexities of life and thought. Without its support man would wander aimlessly through this perplexing universe.

Statistics are the eyes of administration. No statesman can tender sound advice on a problem to his government unless he has adequate statistical data before him to base his judgment upon. Crime, drink-evil, tuberculosis and similar maladies need statistical investigation to suggest remedies for their cure. Budget, a collection of the estimates of revenue and expenditure of a state for the ensuing year, is an unavoidable necessity for an efficient running of government machinery. Its preparation is not possible without statistical records and without their utilisation by a personnel having knowledge of statistical methods. Once the budget is ready, decisions regarding enhancement or decrease in the existing rates of taxation or regarding the exploration of new sources of revenue can be taken up without much trouble.

Statistics are an aid to supervision, particularly in these days of impersonal relationship between the employer and the employee. Every institution, commercial or otherwise, aims at obtaining efficiency with economy. Old plans are substituted by new ones. To test the effectiveness of new policies, the officers must be provided with accurate and concisely tabulated information showing the results obtained.

Statistics are invaluable in business and commerce. To be successful, a producer or a dealer should first estimate the demand for his wares, analyse the possible effects of factors like seasonal variations in demand and changes in taste, fashions and purchasing power of money, and then proceed to adjust his output or purchases to demand. For, if he does not arm himself with a cautious and fairly accurate estimate, he

would either be erring on the side of over-stocking himself and thereby suffering loss, or under-stocking himself and, therefore, losing chances of making profit. Business, indeed, runs on estimates and probabilities. The higher the degree of accuracy of a businessman's estimates, the greater is the success attending on his business. Correct estimation demands a high class of skill which only long experience can ensure. Statistics can be profitably employed in measuring commercial activity. The accounts of a concern undoubtedly show its financial position, but they alone cannot correctly indicate business activity. Statistical averages or indices are, therefore, computed for drawing reliable conclusions. Statistics help the launching of new projects and exploitation of potential markets. In a word, statistics are the life-blood of successful commerce.

Statistics constitute a record of the past knowledge and experience. They help in drawing out standards or 'types' with which results from year to year can be compared, reasons for changes deduced, and effects of such changes on future studied. Experience so ascertained acts as a wise counsel. The businessman, forewarned of a currency inflation, boom or depression, shall prepare himself to face it boldly.

An **underwriter,** a **stock-exchange broker** or an **investor in securities** needs a knowledge of interest rates, the fluctuations of investment market and other data to strike a timely bargain. A **banker,** intending to build up a pyramid of credit, should have an adequate knowledge of the seasonal variations in the calls for money on his bank to decide the amount of reserve that he should keep from time to time in his vaults.

A **railway** operating over a wide area has numerous sources of possible wasteful expenditure. Bad working, such as using four engines where three will do, or handling three tons where four should be disposed off, is costly to the railway itself. Similarly, inability to clearly gauge the necessity of running special trains is not only to cause inconvenience to its patrons, but also to deny to the railway the revenue that could have been its own, if only conclusions had been drawn from past experience based on statistical records. The rod of statistics is, therefore, indispensable for railway company to keep its working within the bounds of efficiency and economy.

All forms of **insurance** subsist on precise calculations based on the analysis of a huge mass of data. The entire working of life assurance schemes, for instance, rests on the compilation of life tables and computation of expectation of life from time to time. Unemployment or sickness insurance similarly depends on statistical data.

Again, in order that a social or economic **legislation** may be fair and judicious it should be based on carefully recorded quantitative information. Enactments with regard to poor-relief, fixation of rate of exchange, levying of excess-profits tax or stoppage of child-marriage need a proper statistical investigation. The merit of the recommendations of various committees and commissions largely rests upon the statistical information behind them and the corrections of the statistical methods used in analyzing them.

Statistics is indispensable in social and economic studies. A sociologist may attempt to demonstrate with the help of statistical data the relationship between sales of liquor and crime, or between suicide and poverty. A theoretical economist may deduce important economic principles from empirical data, or verify the validity of deductive laws of economics by the inductive method of statistics. If one wants to study the march of time as revealed by the trend of population and the production of food, the relation of changes in the value of currency to prices, the relation of railway freights to internal and external trade, the incidence of taxes, the influence of wages on health and efficiency of labour, etc., he must take recourse to statistical investigation. A serious danger in the absence of statistical information is to make arbitrary estimates to suit one's preconceived ideas and pet notions. Statistics brings truth to light and corrects faulty observation. An economist or a sociologist should equip himself with adequate knowledge of statistical methods to guard himself against possible fallacies of argument.

Statistical methods are applicable to all sciences. Astronomy pioneered their use in making predictions about eclipses, and biology has equally appreciably utilized them in its generalizations, for instance, in laws of variations and heredity. Meteorology uses them for weather forecasts. In these sciences and physics, geology, zoology, etc., with whatever care and caution

may the measurements be taken, they cannot always be mathematically exact. And so, an important problem to be tackled in them is to compute the most probable estimate—an average or a type—from a complex group of observations, about which all the measurements are grouped in accordance with some definite law. The next task is to watch the nature and direction of the changes in the type or grouping of the measurements about it. Upon such study are based several generalizations and theories of these sciences.

Scope of Statistical Science.

The above description of the importance of Statistics makes it clear that the science includes in its fold all quantitative analyses, to whatever department of inquiry they might relate. Its scope is, therefore, stretched over all those branches of human knowledge in which a grasp of the significance of large numbers is looked for. Its methods provide an important manner of measuring numerical changes in complex groups and judging collective phenomena. Its scope is thus wide, the limiting factor being its applicability to studies of quantitative character alone.

Limitations of Statistics.

The Science of Statistics is undoubtedly very useful for conducting a scientific inquiry, yet it has certain limitations.

Statistics cannot take cognisance of individual items. It deals with mass phenomena and as such throws light on the characteristics of the *whole* of a given group. Income *per capita* for a country, for instance, is nothing but the summation of its individual incomes divided by the total number of its inhabitants. Individual incomes may differ widely, as they do, from the poor to the rich. But these differences of incomes are brushed off in computing the national average, which gives a graphic idea of the *per capita* income of the *entire* country. If this average is a fairly good figure, it by no means lessens the miseries of the under-fed and ill-clothed paupers and affords no reason why they should not be protected. But the *per capita* income would not reveal the extent of their misery, much less that of the torture of each individual pauper. Consequently, where a study of the individual constituents of a group is

desired, the group average will not yield the required information.

Statistical results should not always be treated as the sole determinants of the value of a group. Statistical method is one of the several methods of studying a problem. It helps us in studying the trends, in framing an idea of the probabilities, in knowing how a given phenomenon has been behaving generally. It does not lead us to unquestionable conclusions. Statistical results are more in the nature of estimates than exact statements. If, however, a statistical inference is confirmed by other available evidences and methods of study, it can be taken as fairly reliable. Notwithstanding this, statistical studies are as necessary for the quantitative appreciation of a problem as accurate measurements are for the construction of a building.

Statistical results might lead to fallacious conclusions if they are quoted shorn of their context or manipulated. If only the average profits earned by two business houses over three years were quoted, and, incidentally, these averages happened to be identical quantities, one might conclude that both the houses were doing equally well. It is not improbable that if one were to glance at the actual profits earned by them during the given period, the profits of one of them might be found to be going up while those of the other going down. Similarly, if the sample selected were biased or data compiled for one purpose were used for an entirely different purpose, or tampered with, fallacious conclusions will naturally be drawn.

Statistical laws are true on the average or in the long run. They are not like the exact laws of physical sciences which are said to hold true in every individual case that is subject to them. Statistical laws, therefore, show approximate tendencies, *e.g.*, Pareto's law of income distribution.

Statistical data must be statistically uniform. That is, they should be homogenous; they should belong to a causal system that is highly stable, so that there shall be no material fluctuations in its main characteristics over the whole field of observation. Without homogeneity of data, comparisons would be vitiated.

Statistical methods are not applicable to the study of those facts that are not quantitatively measurable. Health, culture, character, friendship, skill, poverty, cruelty and pessimism

cannot be quantitatively expressed. In studying any such phenomenon the statistical aspect is assumed to be subsidiary to other considerations. A comparison of the 'state of civilisation' between two countries does not lend itself to statistical treatment. Resort in this case may be had to such numerical data as the number of persons passing a certain standard examination, the number of places of worship or entertainments, or the number of people convicted of crime. But these figures only indirectly relate to the real problem. They are subsidiary to other information like the manner in which people in the two countries live, the value they attach to principles of right conduct, the treatment they mete out to others, the type of work they perform, the food they take, and so on. Therefore, it follows that the use of statistical methods is confined to quantitative studies.

The greatest limitation of statistics is that only one who has an expert knowledge of statistical methods can scientifically handle statistical data, since statistics, like medicines in the hands of quacks, are capable of being easily misused by the inexpert. One might harness statistics to his aid and make the worse appear to be the better case. Many people, therefore, look at statistics with an eye of suspicion.

Distrust of Statistics.

There are said to be three degrees of comparison in lying: lies, damned lies, and statistics. One may not believe in the truth of a statement made without citing statistics. But, when he is presented with figures in its support he is led to believe— 'if figures say so, it can't be otherwise'. Such is the power statistics enjoy. And if this power is misused, say figures are deliberately manipulated, one may, for the time being, be led to take an utterly false statement for an absolutely accurate fact; but truth will be out some day and when it is out figures that were cited in support of the statement would be labelled as lies. Cases, where figures have been put forward as an evidence of the accuracy of a statement otherwise wrong, are not wanting and since lies and damned lies can be detected by a lay mind much more easily than a misuse of statistics, statistics have suffered the stigma of being classed with lies.

But reasons for disrepute cannot lie with statistics. By

themselves they carry no weight. They can support false conclusions just as easily as they support the true ones. With them one may 'prove' that income *per capita* in India is high, while others may 'prove' it to be low. What are these diametrically opposite conclusions due to? They are due either to motive or to ignorance. One may knowingly mis-handle the statistical apparatus with a will to prove his case, or ignorant of the characteristics of statistics and statistical methods run away with any figures, apply to them any technique and feel happy over what he may regard his achievement. Thus it is design and ignorance which lead to misuse or abuse of statistics, which in their turn leads to their disrepute. It is not statistics that are lies. They are only tools in the hands of statisticians. If tools are abused or misused it is not tools which are bad. The fault lies with the way the tools are used.

The popular distrust against statistics is generally expressed in the remark: 'statistics can prove anything'. Those who say so are themselves at fault to a large extent. As a matter of fact, little or nothing can be proved by statistics. What can be done by them is to describe a phenomenon quantitatively, classify it into parts, summarize the facts relating to each part and prepare the ground for a logical inference. Very often too much faith is placed in figures alone and it is believed that the inference to which statistics lead is the *only* inference possible, that the inference is infallible and therefore need not be supplemented or verified by other than statistical evidence. This is what should not be. Conclusions must be made in part on evidence other than that offered by statistics.

If figures are stated without their context, they are applied to a phenomenon other than the one to which they really relate, figures relating to a part of a group are taken as relating to the whole, figures favourable to an argument are stated omitting the other side, they are inaccurately compiled, deliberately manipulated and unscientifically interpreted—in all these cases they can be made to produce a false statistical argument. All these apprehensions make many a man look at statistics with a jaundiced eye.

Statistics suffer from the draw-back that they do not always bear on their face the mark of their quality. To a casual observer, a crude table compiled from unreliable data looks as

valuable as another prepared after great pains by a number of trained statisticians. Most often, people who are served with statistical information do not know whether a particular factor can be statistically evalued, or whether the information is based on satisfactory data. If they are men who believe that 'figures won't lie' they shall accept them without question, while others who are sceptical of their truth shall treat them as 'tissues of falsehood'. In fact, before accepting or rejecting statistics, one should enquire into the competence of their author, or the reliability of the source from which they are obtained.

Another apparent draw-back of statistics is that since they are expressed as definite, concrete quantities they look innocent and precise, and people often believe them to depict an accurate picture. But if they are disillusioned, they blame them. As a matter of fact, their appearance in quantitative form is not a guarantee of their accurately presenting the phenomenon to which they relate. Their reliability will depend on the methods employed in collecting them.

Whatever be the distrust of statistics it does not imply that statistics have no value, or that the Science of Statistics is useless. Drugs may be misused, but neither is their usefulness lost, nor does the science of medicine become valueless for that reason. No doubt statistics do not supply conclusions but they do furnish, in part, the basis on which conclusions may be drawn. Their usefulness is, therefore, great. It is imperative then, that statistical data should be handled only by those who are aware of their use, limitations and dangers and are free from prejudice. If their limitations are forgotten, fallacious conclusions would result. If data are carefully collected and scientifically analysed, the results obtained shall be trustworthy. With the study of Statistics as a science, with the recognition of its limitations and with improvements in its technique, the cause for its distrust is gradually waning. A layman is apprehensive of statistics largely because he does not understand the technique which statisticians apply to a problem in which he is interested. If he learns to exercise his discriminating mind and critical faculty to judge for himself before blindly accepting or summarily rejecting any statistics that he is served with, the misusers of statistics will find it hard to ply their trade.

EXERCISES

(1) Explain and illustrate the functions of statistics.

(2) 'Statistics are like clay of which you can make a God or a Devil, as you please.' Discuss.

(B.Com., Alld., 1948).

(3) 'Statistical methods are most dangerous tools in the hands of the inexpert. Statistics is one of those sciences whose adepts must exercise the self-restraint of an artist.'

Explain fully the significance of the above statement.

(M.A., East Pb., 1951).
(M.A., Patna, 1942).
(B.Com., Alld., 1947).

(4) 'There is scarcely any question in Economics which might not be advanced by bringing to bear on it (*a*) a knowledge of what statistics have to say; combined with (*b*) a knowledge of what statistics can't be made to tell, but which has to be reckoned for in a realistic solution.'

(Marshall).

Illustrate this dictum by reference to an actual economic problem.

(M.Com., Luck., 1944).

(5) 'A knowledge of statistics is like a knowledge of foreign language or of algebra: it may prove of use at any time under any circumstances'—Explain.

(B.Com., Alld., 1946).

(6) Give the important uses and limitations of Statistics. Show its relation to Economics and Mathematics.

(B.Com., Luck., 1938).

(7) Discuss the utility and value of statistics in the measurement of social phenomena. How far do you agree with the view that economic planning without a proper organization of statistics of the country is inconceivable?

(M.A., Agra, 1940).

(8) 'The Statistics of a business can be treated scientifically and the preparation and study of business statistics may be made a more exact science than the study of national and social statistics.' Explain.

(B.Com., Alld., 1932).

(9) Discuss the role of Statistics in the scientific investigation of phenomena. What are its limitations?

(M.A., Alld., 1951).

(10) 'The application of statistical methods to investigations in the technological (and indeed any other) field is based on assumptions, is subject to limitations, and often leads to uncertain inferences'. (Tippett)—Discuss.

(M.A., Alld., 1952).

(11) "There are three degrees of comparison in lying: There are lies, there are damned lies, and there are statistics"— How far do you agree with this statement?

(12) How do you reconcile the following pairs of statements?—

I. (*i*) Statistics can prove anything.
(*ii*) Statistics do not prove anything.

II. (*i*) Statistics are lies.
(*ii*) Figures do not lie.

(13) Describe briefly the different kinds of statistical methods, and show the usefulness of the knowledge of statistics to businessmen, scientists, and economists.

(B. Com., Agra, 1936).

(14) Explain clearly what *you* understand by the Science of Statistics. Discuss its scope and limitations.

(B. Com., Alld., 1944).

(15) 'Sciences without statistics bear no fruit, statistics without sciences have no root.'

Explain the above statement with necessary comments.

(M.A., Patna, 1943).

(16) In what ways can statistical methods be misused by interested persons? Give at least two examples of the misuse of statistics.

(B. Com., Luck., 1939).

(17) Discuss the importance of the study of statistics, and show how it can help the extension of scientific knowledge, the establishment of a sound business, and the introduction of social and political reforms.

(B. Com., Agra, 1942).

(18) 'Statistics only furnish a tool, necessary though imperfect, which is dangerous in the hands of those who do not know its use and deficiencies.' (Bowley)—Discuss.

CHAPTER IV

PLANNING A STATISTICAL INQUIRY

A statistical inquiry passes through several stages from its planning to the writing out of the actual report. These stages may be conveniently grouped as under:

(1) Planning the inquiry;
(2) Collection of data;
(3) Scrutiny of collected data;
(4) Classification and tabulation of data;
(5) Presentation and comparison of data through charts and graphs;
(6) Analysis of data through computation of derivatives, coefficients, etc.; and
(7) Interpretation of data and writing up the report.

Generally speaking, organisation of a statistical inquiry covers the first two stages mentioned above. In this Chapter, only the first stage is discussed; the rest of the stages will be taken up in subsequent Chapters.

Facts and figures constitute the foundation of all statistical inquiries. Therefore, actual collection must be preceded by a well-thought out plan. In planning a statistical inquiry, it is necessary to

(1) ascertain the 'object and scope' of the inquiry;
(2) determine its 'nature and type';
(3) define 'statistical units'; and
(4) lay down the 'degree of accuracy' desired.

Object and scope of inquiry.

Every inquiry has an object and a certain scope. It is essential to determine them clearly and unambiguously, for then alone can the organiser of the inquiry decide what part of

his investigation can rest on a statistical basis and for what part he will have to rely on other than statistical evidence. Determination of the object and scope is also an essential preliminary to preparing the actual plan of investigation.

Suppose food controls are relaxed in a statutorily rationed town so that the fair-price shops run by the Government to sell limited quantities of foodgrains at fixed prices and the open market both function simultaneously in that town, and it is desired to study the effect of this relaxation on people's purchases. Here the object and the scope of the inquiry should first be precisely defined, making it clear as to whether it is intended to study the effect on people's purchases from the fair-price shops only or in the open market also; what period after the relaxation of controls is the inquiry intended to cover; whether the inquiry will extend to the entire town including its suburbs if they were formerly under statutory rationing or whether the municipal limits only will be covered; whether all classes of people will be studied or the inquiry will be confined to people of a certain income-group only. Firm figures about the purchases from the fair-price shops will no doubt be available from the records of the fair-price shopkeepers and the rationing offices. If information about the quantities being purchased from the open market is also to be collected, inquiries will have to be made of the purchasers. So far, the investigation will have a statistical basis; but if it is also desired to know as to why the people are purchasing from the open market also in spite of, let us say, the prices in the open market being higher than those in the fair-price shops, or why they have reduced their purchases from the fair-price shops, this part of the inquiry will assume the nature of a public opinion survey and may not be found to possess the same degree of reliability as the statistical part. The two parts of the inquiry will have to be kept separate and distinct.

If the object and scope of the inquiry are not clearly determined beforehand, difficulties will arise as the inquiry proceeds, which will lead to confusion and waste of time and money; it may even happen that unimportant and useless data may be collected in greater detail at the cost of the information really required for studying the given problem.

Nature and type of inquiries.

Statistical inquiries are of several types. Broadly, their nature may be distinguished as follows:—

(1) **Official, semi-official or non-official.**—An official inquiry is conducted by or on behalf of a State Government or the Central Government; a semi-official inquiry by such bodies as enjoy government patronage and non-official inquiries by private bodies or individuals. The facilities that these three bodies have for the collection of data differ. The government may exercise 'compulsion' through legislation, a semi-official body may 'request' while a non-official body may have to 'beg' for the purpose. The government can spend much more money on an inquiry than a semi-official or non-official body. Also, facts material to an official or a semi-official inquiry may not carry the same emphasis in a non-official inquiry.

(2) **Primary or secondary.**—In a primary inquiry original data are to be collected, while in a secondary inquiry data are already available in office records or published documents and have to be culled from them. For example, census of population yields the figures which it is the purpose of the census to collect; collection of data is the sole end here. On the other hand, if one were to study the population problem of the country and to cull out relevant data from the census tables, analyse and interpret them, it would constitute a secondary inquiry. In the case of a primary inquiry, the definitions of the units and the scope of the inquiry can be made to suit the object of the inquiry in hand; in the case of the secondary inquiry, one will have to go by the definitions already used in the collection of data available in the relevant documents. In a primary inquiry regarding the need for dearness allowance in times of rising prices, data about 'real wage' may be collected. If, on the other hand, this were conducted as a secondary inquiry, one will have to depend on published data about wages which are in terms of 'money wages' generally.

(3) **Census or sample.**—A census inquiry implies complete enumeration of each unit of the universe while a sample inquiry connotes the study of the universe on the basis of only a representative part of it. Population Census in India is an example of complete enumeration while the Agricultural Labour Inquiry that of a sample survey.

(4) **Mail-card inquiry or inquiry through investigators.**—The mail-card inquiry is one in which the informant is given a list of questions or a blank form which he returns to the sender by post or by hand with his answers duly filled in. For administering a scheme of food rationing in a town, every householder may be made responsible for supplying information relating to the number and age of persons in his household. Since the information in this case will have to be supplied by persons of varying educational standards, the questions contained in the form should be few, simple and unambiguous. The scope of the inquiry would naturally be limited on this account. If, on the other hand, skilled investigators are appointed to approach the informants and collect requisite information from them, the scope of the inquiry can be enlarged so as to include a larger number of questions, some of which may even be complicated ones. The mail-card inquiry is less expensive but the response from the informants may not be much unless the furnishing of returns is made compulsory or is in the interest of the informants themselves.

(5) **Initial or repetitive.**—An initial inquiry is one which is carried out for the first time. For this inquiry, a plan will have to be formulated. A repetitive inquiry is conducted in continuation or repetition of previous inquiries. In this case, the old plan with such modifications as experience or necessity demands may be adopted, care being taken to see that the modifications should not materially affect the comparability of information collected during the previous inquiries.

(6) **Direct or indirect.**—In a direct inquiry, data pertaining directly to the subject-matter of the inquiry are collected. Where the subject-matter of the inquiry is not capable of being statistically evalued, some allied information reducible to numerical quantities and bearing indirectly on the subject-matter may have to be collected. This is an indirect inquiry. Height of students in a class is measurable directly in feet and inches but their intelligence cannot be quantitatively ascertained. In the latter case, reliance may have to be placed on indirect evidences of intelligence, as for example, the marks obtained at a certain examination, or intelligence tests.

(7) **Confidential or non-confidential.**—Confidential inquiries are those which are conducted for official use only and are not

thrown open to public. Trade associations may, for instance, collect information from their members and keep such information secret. Such inquiries are generally conducted on a restricted scale and with great care. Non-confidential inquiries, the results of which are not kept from the public, may have a wider scope.

(8) **Regular or ad hoc.**—A regular inquiry implies collection of data on a regular periodical basis over a period of time, while an *ad hoc* inquiry is conducted once in a while. Prices of different commodities may, for instance, be collected on a weekly basis for years on, for constructing an index number series; but an inquiry into the extent of rural indebtedness in the country may be made only casually.

The above classification of statistical inquiries into various types is only intended to afford a convenient picture of their differing nature; the classes are by no means rigid or mutually exclusive. The Agricultural Labour Inquiry, 1950, was, for instance, an *official* inquiry, conducted for collecting *primary* data on wages, cost of living, etc. of agricultural labour, organised on random *sampling* basis, wherein data were collected *through the investigators*. It was an *initial* inquiry and dealt with the subject-matter *direct*. It was *non-confidential* and *ad hoc*.

Statistical Units.

Quantitative science demands a precise and unambiguous terminology in respect of the units in which the aggregates to be counted shall be expressed. It also demands strict adherence to that terminology throughout the inquiry. The units of measurement should therefore be defined, rigidly and unmistakably, before the inquiry starts and this definition should be stuck to during the course of the inquiry. This will ensure uniformity in the data collected, which is necessary for effecting comparisons and drawing valid inferences.

The task of defining the statistical units seems at first easy but in many cases the opposite turns out to be ultimately true. 'Area under cotton' appears to be a simple expression; but while defining it, several questions arise: Is it the area sown to the crop or the area from which harvest was reaped? Should it include the area occupied by boundaries between the

adjoining fields or by bunds, embankments and channels? If cotton is sown mixed with groundnut, how should the area under the two crops be separately assessed? Similarly, what does the term 'wholesale price of goods' signify? Does it stand for the price at which the producer sells the goods concerned to the stockist or the one at which the latter sells to a wholesaler? Is it the price at which the market opened on the day of enumeration or the one at which it closed? Should it include the cost of bags or containers, handling and weighment charges or be exclusive of such incidentals? It will be seen that the correct definition of a statistical unit will be determined by the object of the inquiry and that different objects will necessitate different definitions of the same unit.

A statistical unit should satisfy the following requirements:—

1. The unit must suit the purpose of the inquiry.
2. Its definition must be unambiguous, simple and complete in itself.
3. The unit must be definite, specified and ascertainable.
4. It must be stable and standard. (In India, currency fluctuations have not been rare, and weights and measures still vary from region to region. Hence the necessity of taking stable and standard unit.)
5. It must ensure homogeneity and uniformity. It should not connote different characteristics at different times during the same inquiry. Data which are not homogeneous cannot lead to valid comparisons. The golden rule of comparison is that the like alone can be compared with the like. If the data are heterogeneous, they may be broken up into small homogeneous classes or resort may be had to the method of standardization to convert them to a uniform basis. For example, the data for the compilation of an average of the wages received by workers in a factory, where adults and children of both the sexes are working side by side, are heterogeneous. These data many either be sub-divided into three groups, *viz.*, 'wages for male adults', 'wages for female adults' and 'wages for children', or females and children may be expressed in terms of equivalent men, before striking their average.

The units of measurement may be classified into:

I. Units of Enumeration, and
II. Units of Analysis and Interpretation.

Units of Enumeration: Simple and Composite.—Units of enumeration are those in terms of which measurements are made, or data collected. These units are of two types: simple and composite.

A **simple unit** is one that denotes a combination of characteristics that occur together. It simply distinguishes classes. Such units are mutually exclusive. They are defined easily and fairly precisely. The degree of error associated with them is, therefore, small. Examples of simple units are a ton, a passenger, an accident, a sale, a store, a house, a room.

A **composite unit** is one that is formed by adding a limiting or qualifying word or phrase to a simple unit, with the result that its scope becomes limited and the task of defining it does not remain easy. Examples of composite units are ton-mile, passenger-mile, industrial-accident, credit-sale, chain-store, bond-house, bed-room.

A simple unit, like accident, is a general term. The cause of the accident does not matter here. In includes industrial-accident, train-accident, fall from the roof of a house, etc. A composite unit, such as an industrial-accident, is not so general; it is limited to a specific type of accident and excludes accidents due to other causes.

Units of Analysis and Interpretation.—These units are those in terms of which data are compared. They include rates, ratios and coefficients. (They will be discussed in detail in subsequent Chapters.) A **co-efficient** takes the form of a comparative statement. It may relate to time (as, for example, wages expressed in Rupees per month) or to space (such as, yield of wheat per acre) or to conditions in time or space (say, income *per capita per annum*). A co-efficient, in effect, is a comparison between the numerator and the denominator, both of which should be related and homogeneous. If passenger-miles are divided by passenger-train-miles, we shall obtain passengers per train. But if passenger-miles are divided by ton-miles, a monster will result.

Degree of Accuracy.

The last step in planning a statistical inquiry is to lay down the standard of accuracy to which the data to be collected

should conform. It should, for instance, be decided whether prices or wages shall be quoted correct to a pie, an anna, or a Rupee. The degree of accuracy required will vary with the object of the inquiry. The aim should be to determine it in advance, clearly and precisely, so as to avoid the necessity of conducting a second inquiry. Time, labour and money will thus be saved from being wasted.

The next stage in organising an inquiry is to decide on the method of collecting statistical data.

EXERCISES

(1) Describe the preliminary steps you would take in planning a statistical inquiry.

(2) What do you understand by the 'object' of an enquiry? Why is it necessary to determine it before planning a statistical inquiry?

(3) What are statistical units of enumeration? How do they differ from the units of comparison?

(4) Differentiate between simple and composite units. Give illustrations of transforming units from simple to composite.

(5) What is a statistical unit? Is it necessary that the data be homogeneous?

(B.Com., Agra, 1939).

(6) Describe the various stages in conducting a primary economic investigation. What precautions will you take at each stage?

(M.A., East Pb., 1950).

(7) By the use of statistical methods we try to channel and appraise the flood of data relating to business operations or social processes that the fact-gathering-agencies and Government release upon us.

Discuss the statement and describe briefly the work of a statistician from the organisation of data to the presentation of a statistical report.

(M.A., Patna, 1944).

CHAPTER V

COLLECTION OF DATA

Primary and Secondary Data.

Statistical data are generally classified as primary and secondary. The former are those which form the raw material of inquiry, while the latter are those which have gone through the statistical machine at least once. The former are original, *i.e.*, those in which instances have been recorded as they occurred without having been grounded at all. The latter are those that have been worked up to a certain extent, *i.e.*, they have been tabulated and presented in some suitable form for a certain purpose. They are generally expressed in totals, averages and percentages.

Distinction between primary and secondary data is one of degree. Data which are secondary in the hands of one may become primary in those of another. For instance, statistics of foreign trade of India published in the *Monthly Accounts relating to the Foreign Sea, Land and Air-borne Trade and Navigation of India* are secondary data to the general public, while they are primary data in the hands of the statisticians of the Department of Commercial Intelligence and Statistics who compile the publication.

On the basis of primary and secondary data, the methods of collecting statistical material have been divided into Primary Method and Secondary Method.

Primary Method.

Under this method are grouped the following methods of collecting data:

1. **Direct Personal Observation.**—This method implies that one who has to collect the data should personally study the concerned phenomenon and obtain first-hand information. If

an inquiry is to be conducted into the family budgets and living conditions of industrial labour, the investigator should himself live in the industrial area as one of the industrial workers, mix with other residents and make patient and careful personal observation regarding how they spend, work and live. Information thus collected will be reliable. But, since within a reasonable amount of time an extensive field of inquiry cannot be covered by this method, it is useful specially for intensive and localized inquiries. This method is also open to the charge that the chances of personal prejudices of the investigator affecting his conclusions, even though unconsciously, are great.

If it is not practicable for the investigator to be on the spot or to devote the time the above method demands, an alternative for him is to question and cross-examine the persons who are directly in touch with the facts under investigation. Here, since the investigator counts on the goodwill of others, he will have to be courteous in his behaviour. Besides, the questions that he would ask must be few, very simple, clearly worded, not inquisitorial and so far as possible demanding an answer in 'yes' or 'no' or 'a number'. This alternative method is also used in intensive studies.

2. **Indirect Oral Investigation** assisted by a standard list of questions.—When the information desired is complex and informants are indifferent to supply it if directly approached, or if the field to be covered is very extensive so that the first method cannot be successfully employed, the *viva-voce* indirect evidence of several third parties, preferably of those indirectly in touch with the facts under inquiry, may be recorded. Commissions and Enquiry Committees appointed by governments generally find this method suited to their needs. But certain precautions must be observed in order that reliance may be placed on the data so collected. Firstly, the indirect evidence of one person should not be relied upon. Secondly, it should be clearly ascertained whether the informant really possesses a knowledge of the full facts. Thirdly, it should be considered whether the person questioned is prejudiced in favour of or against a particular viewpoint or is motivated to colour the facts. Due allowance must be made for the optimism and pessimism of the informant. Lastly, if the informant happens to be an uneducated person or suffering from occasional fits of

mental disequilibrium, it should be seen whether he is in a position to give expression to his ideas adequately and precisely at the time he is questioned.

3. **Estimates from Local Sources or Correspondents.**—This method does not imply a formal collection of data. Local correspondents obtain the estimates in their own manner and report them to an appointed authority. This method yields only approximate results, but expeditiously, at a small cost and with ease. For instance, estimates of agricultural wages may be periodically furnished to the government by village school teachers.

4. **Investigation through schedules to be filled by the informants.**—This method differs from the preceding one in that the questions asked of the informants are those in respect of which they are supposed to have definite and precise information. If the informants reply intelligently, this method is good for extensive inquiries. It is inexpensive and fairly expeditious. This plan is largely adopted by private individuals and even the government. But a large number of informants do not generally answer the schedules unless the inquiry is in the informants' own interest, or the private individual or institution responsible for the inquiry is able to persuade them to answer, or the state exercises compulsion through legislation. And, schedules that are returned are very often incomplete, ambiguous and full of errors, since the average informant is indifferent in these matters. In order that correct answers may be had, the schedule, or a letter of request attached to it, should state the purpose of the inquiry, the identity of the person or the institution responsible for the inquiry and an assurance to treat the information tendered as confidential, if so desired by the informant. The questions should be few, clearly phrased and, above all, simple, and should not be such as to arouse suspicion and prejudice in the informant.

5. **Investigation through schedules in charge of Enumerators.**—Under this plan the informants themselves are not to fill in the schedules. Instead, the trained enumerators put them questions and record their answers. Therefore, in this kind of inquiry the schedules can be much more exhaustive than those in the previous one, and the scope of inquiry can also be enlarged. Correct interpretation of every question and the

method of collecting information must be explained in detail to the enumerators, so that different enumerators may not give different weight or meaning to the same question. Enumerators should be equipped with a sample schedule duly filled in. *This plan affords quite good results and is the best for many extensive investigations.* It is generally adopted in large-scale governmental inquiries. It was, for instance, employed in conducting the Agricultural Labour Enquiry organised by the Government of India in 1949-50. Its cost is, no doubt, prohibitive to a private individual or institution.

Where investigations are carried out with the help of enumerators, great care should be exercised in selecting the proper personnel and giving it adequate training.

Choice of Enumerators.

The quality of the investigation depends on the enumerators. Intelligence, diligence and integrity must be their attributes in order that vague replies of informants may be detected, corrected or eliminated and fictitious quantities may not be entered. The enumerators should not tend to be too officious; they should, on the other hand, be courteous and tactful so that they may extract the requisite information without causing resentment or ill-will in the informants. They should be free from bias. They should be conversant with local customs, manners and language. If these precautions are taken in the selection of enumerators, needless errors and confusion shall be avoided.

Training of Enumerators.

The enumerators should be given adequate training before being sent out to do field work. First, the implications and scope of each question to be put to the informants should be explained to the enumerators. They should be instructed as to how to check up evasive and apparently wrong replies without being argumentative with the informants. Then, they should either be attached to experienced enumerators for some time to acquire practical training or taken to a pilot inquiry if one is conducted before starting the full inquiry.

Choice of Questions.

A word about schedules used is necessary. The schedules

may be either what are called 'Questionnaires,' the answers to which are recorded on a separate piece of paper, or 'Blank Forms,' in which space is provided in the form itself for filling in the reply. Headings and titles drawn up in them should be lucid and easily understandable, and the degree to which accuracy of a numerical result is required should be indicated. The size of the paper should not be unwieldy, and each word and phrase used should be carefully scrutinized for ambiguous or controversial interpretation.

The questions which are asked should be—

(1) such as the informant shall be able to answer;
(2) few in number;
(3) simple and clear enough to be easily grasped;
(4) not inquisitorial and not causing resentment, so far as possible;
(5) requiring brief answer, say 'yes', 'no' or 'a number';
(6) corroboratory, if possible;
(7) capable of being answered without prejudice; and
(8) directly related to the point of information desired.

Selection of Representative Data.

When the inquiry in hand is very extensive, it will not be practicable to undertake a census type of inquiry where each individual item of the universe or 'population' shall be questioned. The inquiry will have to be of the type of sample survey, and the sample will have to be representative of the whole field.

The methods of selecting a sample are two:—

1. Deliberate or Purposive Selection,
2. Random Sampling or Chance Selection.

Deliberate Selection.—In adopting this method the organisers of the inquiry deliberately choose the particular units on the basis that the small mass that they so select out of a huge one will be typical or representative of the whole. If economic conditions of people living in a State are to be studied according to this method, a few towns and villages may be purposively selected for intensive study on the principle that they shall be representative of the entire State. In this method personal element has a great chance of entering into the selection of the sample. The investigator may select a sample which shall yield results favourable to his point of view and the entire inquiry

may thus be vitiated. If the investigators are unbiassed, the results obtained from an analysis of deliberately selected sample may be tolerably reliable, provided the basis of selection is unquestionable from other points of view.

Random Sampling.—Chance selection, on the other hand, is free from the influence of the personal factor. It is, so to say, a lottery method in which individual units are picked up from the whole group not deliberately, but by some mechanical process, so that every unit has equal probability of entering the sample. Here it is blind chance alone that determines whether the one unit or another is selected. This is an important condition of this method. If from a list of villages of a given area, arranged in alphabetical order, every 100th or 50th village is marked out for intensive study it would give a hundredth or a fiftieth sample of the whole area. But, once a random selection of villages has been made, on no account should any of the selected villages be substituted by another.

Census vs. Sample.—In a census inferences for the whole population are drawn from a study of the whole field, while in a sample survey inferences for the whole population are drawn from the study of a representative part. The 'Sample' method has many advantages over the 'Census' method. It affords a sufficiently accurate picture of a large group without the need for a complete enumeration of all the units of the group. It economises in time, labour and money and permits a small band of skilled investigators to do the whole job more efficiently and precisely than a large army of unwilling, inefficient enumerators would do in a census. Random sample survey is largely coming into use because of these advantages and also because of its being more scientific than 'purposive' sample.

Theory of Probability.—The method of random sampling is based on the mathematical 'Theory of Probability'. The theory implies that if from a very large group of items, technically called the 'population', a moderately large number of items is chosen *at random* such numbers are almost sure, *on the whole,* to possess the characteristics of the population. If two men randomly plucked 200 leaves each from a particular tree, the averages of the lengths of the leaves plucked by each man would be almost identical. Further, if one were to obtain the average length of all the leaves of the tree, it would not

materially differ from the average length of either group. Similarly, if a rupee coin is tossed twelve times, the probability is that it will fall half times (*i.e.*, six times) with its head or tail turned upwards. On this principle gamblers—dice-throwers, card-players, etc.,—run risks continuously and with profit, and insurance companies insure against death or other calamities. It is this principle that is responsible for regularity in the numbers of crimes and of suicides in a country for a given period of time. This principle is christened the **Law of Statistical Regularity.**

It should, however, be noted that *any* number of samples will not give exactly the same results as a study of the whole group would. As a matter of fact, the probability of error diminishes with an increase in the number of items included in the sample. That is, the larger the sample, the more reliable are its indications.

Law of Inertia of Large Numbers.—This law furnishes a corollary to the law of statistical regularity. It implies that large numbers are relatively more stable than small ones. If one part of a large group varies in one direction, the probability is that another equal part of the same group would vary in the opposite direction so that the total change would be slight. For example, the production of tea or wheat may vary from locality to locality in a given year, but the total world production of tea or wheat remains relatively stable for decades. Thus large numbers have great inertia.

But it does not mean that the property of inertia does not allow for change with the passage of time. It only signifies that if the numbers under consideration were of great magnitude, the change would be more regular than in cases where small quantities were considered. Secular movements resulting from long period tendencies in the background of conditions are not precluded. The death-rate of a country may for instance be relatively stable from year to year, but for a long period it may show a progressive decline.

Ensuring collection of reliable data.

Collection of reliable data is a prerequisite to drawing trustworthy conclusions from the inquiry. Unreliable data cannot lead to unquestionable inferences. Where the data are collected by a number of enumerators, it is a cooperative effort.

It will have to be ensured that this cooperative effort results in the same quality of data as would be obtained if one and the same individual were to conduct the entire inquiry. Every precaution should therefore be taken to see that the data collected are reliable. These precautions include the following:—

(1) The object of the inquiry should be precisely determined.

(2) Statistical units should be rigidly defined and the definition once specified should be adhered to throughout the inquiry.

(3) The degree of accuracy desired in the data should be laid down in advance.

(4) There should be no room for the bias of the enumerator, the informant and the measuring instrument to enter the data. This is nowadays sought to be achieved by resorting to random sampling in the case of sample surveys.

(5) The enumerators should be furnished with detailed instructions about the meaning and scope of each question and the manner of filling up the returns.

(6) Proper training in the art of conducting the investigation should be given to the enumerators before they proceed with the inquiry.

(7) In the questionnaires or blank forms, provision should be made for interchecking so that the correctness of one statement may be checked up with reference to another allied statement made elsewhere in the reply to the questionnaire or in the blank form.

(8) An enumerator should not preferably be kept in one place for too long. The enumerators may collect the data in several rounds such that if one enumerator covers questions No. 1 to 10 in a village, another enumerator may visit the same village later on to cover questions No. say, 8 to 17, so that questions No. 8, 9 and 10 will be common to both the enumerators. Thus, a test-check of the accuracy with which the first enumerator filled up the returns will be made automatically.

(9) The collection of data and their analysis should not preferably be entrusted to the same person.

(10) As periodical returns begin pouring in from the enumerators, immediate scrutiny thereof should be taken up at

the headquarters. In case it is discovered that some investigators have not corectly understood the meaning or scope of a certain question, immediate instructions on the particular issue should be sent out to the investigators.

Secondary Method.

Secondary method implies collection of secondary data, that is, data which have already been compiled by government, semi-government or non-official bodies and are available in office records, publications, reports, books, journals, etc. The main sources of such data are:—

(1) Official publications brought out by the Central and State Governments;
(2) Reports published by municipalities, district boards, etc.;
(3) Reports of Committees and Commissions;
(4) Publications brought out by research institutions, boards of economic inquiry, universities, etc.;
(5) Technical, economic and commercial journals;
(6) Publications of trade associations, millowners' associations, Chambers of Commerce, etc.;
(7) Market reports and trade bulletins issued by stock exchanges, large business houses, etc.;
(8) Publications brought out by individual research workers.

The above are illustrations of published data. Secondary data may also be available in unpublished or manuscript form. They may, for instance, be available in the records of government offices or business houses, such as account books. They may also comprise of unpublished notes of investigators and chroniclers.

Precautions in using Secondary Data.

Secondary data should be used only after careful and searching inquiry and due criticism. It is never safe to take them at their face value without fully ascertaining their meaning and limitations. First, one should find out the name of the organisation by whom the data have been compiled. Then, an assessment should be made of the reliability of the compiler and of his ability to procure correct statistics. If the compiler's

standing is not well-known, one may study such parts of the data as are familiar to him and check if they appear to be reasonably correct. In case compulsion was exercised by the government, the informants might have given information reluctantly and, no wonder, not precisely. It may not be appropriate to utilise such information. Similarly, questions relating to use of intoxicants, mental infirmity, total earnings, profits, etc., are sometimes inadequately answered since they arouse suspicion among the informants. Such data also are of doubtful value. In short, one should see whether the details compiled by another organisation or compiler tolerably agree with the details that one would have himself collected if he had personally conducted the inquiry. If there is reason to believe that the compiler was biassed or motivated, the data should be rejected at once. If, on the other hand, the reliability and the competence of the compiling organisation are not doubted, the data may be used but after inquiring further into the following points:—

(1) The scope and object of the inquiry conducted by the original compiler.
(2) The definition of units in which the data are expressed.
(3) The sources of the compiler's information.
(4) The method of his collection, including instructions given to the enumerators.
(5) The degree of accuracy desired and achieved.
(6) The extent to which the data refer to homogeneous conditions.
(7) The extent to which they appear to be comparable over a period of time.
(8) The suitability of their application to the given problem.

If after making the above inquiries, it is established beyond doubt that the data are satisfactory, they may be culled out from the relevant sources and utilised for the purpose of the inquiry in hand.

EXERCISES

(1) How will you organize an economic survey of a

small Indian State comprising five towns and one thousand villages?

(M.Com., Alld., 1943).

(2) How far do the results of statistical investigation depend upon correct sampling? Compare the different methods used to secure representative data.

(B.Com., Agra, 1939).

(3) 'In collection of statistical data commonsense is the chief requisite and experience the chief teacher.'

Discuss the above statement with comments.

(M.A., Patna, 1941).
(B.Com., Luck., 1940).
(B.Com., Alld., 1939).

(4) Compare the advantages and disadvantages of the 'Census' method (or complete enumeration) and the 'Sample' method of collecting statistics.

(B.Com., Cal., 1937).

(5) Mention the different kinds of statistical methods generally used in investigations. Are there any fields of inquiry where these methods cannot be used satisfactorily?

(B.Com., Agra, 1940).

(6) What is a Questionnaire? How does it differ from a Blank Form? What precautions should be taken in drafting questionnaire?

(7) How would you organise an enquiry into the cost of living of the student community in Amritsar? Draw up a blank form or forms to obtain the required information.

(M.A., East Punjab, 1951).

(8) How would you organise an investigation into the handloom weaving industry of the U. P.? Prepare questionnaire suitable for the purpose.

(B.Com., Alld., 1942).

(9) 'Though figures cannot lie, yet liars can figure.'

Expand the above statement so as to explain its bearing on the use of secondary statistical data.

(M.Com., Alld., 1945).

(10) Show the necessity of the use of the method of

random sampling in any extensive investigation. How would you make use of this method in carrying out an economic survey of the rural areas of U.P.?

(B.Com., Alld., 1935).

(11) Statistical investigations carried out by the Government are usually based either on complete enumeration of the universe of reference as, for instance, the population census, or on the study of 'typical' cases as, for instance, the proposals regarding the economic census. Explain why the method of random samples is to be preferred to either of these methods.

(M.A., Alld., 1935).

(12) Write brief, but lucid, notes on:

(*a*) Law of Statistical Regularity,
(*b*) Law of Inertia of Large numbers,
(*c*) Primary and Secondary methods of collecting data.

(13) Classify the methods generally employed in the collection of statistical data, and state briefly their respective merits and demerits.

(B.Com., Alld., 1946).

(14) If you are appointed to investigate the housing conditions of industrial labour in Lucknow how would you proceed to do the job? Give a specimen of the questions that you would put.

(B. Com., Luck., 1944).

(15) 'Fecundity among the poor is higher than among the rich.' This statement is often made. It means that poorer classes of persons beget more children than richer ones. How would you verify statistically the truth or otherwise of this statement? Describe fully the method you would employ, giving examples.

(B.Com., Agra, 1934).

(16) Draw up a suitable questionnaire for surveying the economic aspects of any Cottage industry in which you may be interested. Briefly indicate how you would proceed to collect the relevant materials.

(B.Com., Luck., 1943).

(17) "It is never safe to take published statistics at their face value without knowing their meaning and limitations, and it is always necessary to criticize arguments that can be based on them." (Bowley)—Elucidate.

(B.Com., Alld., 1946).

(18) A sugar manufacturer in the U. P. is anxious to find new markets for his sugar outside India. Describe the procedure he should follow to get all the necessary statistical information for the success of his mission.

(B.Com., Luck., 1938).

(19) What precautions would you take to ensure that the data collected by a number of enumerators are reliable?

(20) What points should be taken into consideration while selecting the enumerators and choosing the questions to be put to informants?

CHAPTER VI

EDITING THE COLLECTED DATA

Editing Primary Data.

After the schedules have been returned by the informants or enumerators, as the case may be, they should be edited, *i.e.*, scrutinized to detect errors, omissions and inconsistencies. Definite procedure for scrutiny of returns should be laid down so that all those concerned with scrutiny do their job on uniform basis. These rules would, for instance, say that the summation of this and that column should equal the figures given in another column; that the sum of monthly figures should equal the annual total, and so on. If possible, defective schedules should be sent back for correction, or, if the investigator has reasonable and strong grounds to do so, he may himself make the required amendments. Undue tampering with them is, however, dangerous. Only in cases of unmistakable error should alterations or modifications be made. Such schedules as are thoroughly unsatisfactory must be rejected. For, a smaller number of correct samples is better than a large number of incorrect ones. In the former case, the error can be mathematically corrected with approximate accuracy. It is not possible to do so in the second. If the majority of informants have misunderstood a question there is a clear case for making a change and conducting a second enquiry. The extent to which omissions may be allowed is also of importance. If the returns unquestionably confirm a certain fact and the samples are tolerably representative, the omission of a number of returns does not matter. If, on the other hand, evidence is conflicting, the omission of even one return may be a serious matter. The degree to which lack of accuracy, or approximations, or presence of errors is to be tolerated in editing the data is of great significance.

Accuracy.

Perfect accuracy in the data is rarely attainable. Wheat

crop in India, for instance, cannot be exactly measured. It can be estimated to a reasonable degree of accuracy. A weighman, however perfect his weighing balance, cannot weigh a commodity correct to within, say, 1/64th of a seer. Similarly, the distance for a four-mile cross-country race cannot be measured without giving a probable error of a few yards. Even in scientific measurements absolute accuracy is unattainable, for, heights of liquids in test-tubes may vary by a thousandth part of an inch or angles may differ by a hundredth part of a degree. Thus, absolute exactitude is not possible; a closer approach to it is. Reasons for it are obvious: the observer and the observing instruments are both sources of error; statistical data cannot be given hard and fast definitions; the statistician, unlike a chemist, cannot experiment, conditions being outside his control. Statistical methods do not, therefore, aim at arithmetical precision. A statistician would be satisfied with *reasonably* accurate figures provided he can measure their reliability. To attempt to obtain the greatest possible degree of accuracy is also to waste time where large figures are involved. At any rate, where notional figures would do meticulous accuracy is superfluous. Statistics has, thus, to deal with estimates and probabilities and not exact enumerations.

What is a reasonable degree of accuracy shall be determined by the nature of the material and the purpose of its measurement. In common use, only a certain conventional accuracy is required. Precious metals or drugs are much more minutely weighed than hay, husk or saw-dust. Height of a room may be measured correct to an inch, but the difference of half an inch in the length of a man's nose will make him a monster. A railway is satisfied if a parcel for booking is weighed correct to a seer, but the post-office would weigh it correct to a tola at least. It would be enough if we can measure an object to that degree of accuracy which is required for practical purposes. Thus **it is relative and not absolute accuracy that is desired in statistical data.**

Approximation.

Numbers, each containing numerous digits, confuse the mind. For convenience, they may be rounded off. For example,

the figure 264,571 may be expressed as 264,600 or 265 thousands, or the population of India for 1951 may be expressed as 361 million or 36 crores instead of 361,101,760 persons.

In approximation all figures except the first digit beyond the margin of accuracy should be left out. For instance, if the length of a leaf is recorded as 3·97 centimeters, though with the scale used it was possible to read correctly to the nearest tenth of a centimeter, it is better to retain the final digit 7, since 3·97 is a closer approximation to the real length than 4·0 cms., which would be the reading if the digit 7 were dropped.

Again *if the last correct figure is a cipher it must be so entered.* If a leaf measures very nearly three centimeters, it should be expressed as 3·0 cms., and not simply as 3 cms., for the latter figure might mean that the reading is accurate to centimeters whereas the former indicates that the reading is correct to millimeters which really is the case in our example. If the reading is correct to centimeters, any leaf between 2·5 and 3·5 cms. in length would be entered as 3 cms., while if the reading is correct to millimeters the entry 3 cms. would mean that the length is between 2·95 and 3·05 cms. Similarly, an entry of 3·00 will mean that the reading is accurate to hundredths of a centimeter, and that the length of the leaf is between 2·995 and 3·005 cms.

When certain digits of a number are to be dropped, all *fractions over half should be counted as whole numbers and all under half discarded.* Fractions equalling exactly one-half may be allowed to remain or left out at discretion. The following table giving approximations should be carefully studied:

Original Number.	Approximation correct to one decimal place.
15·049	15·0
15·050	15·1
15·249	15·2
15·257	15·3
15·948	15·9
15·951	16·0

Statistical Errors.

The word 'error' is used in a special sense in Statistics. It

denotes the difference between the true value and the estimate of a quantity. It shows by how much an estimate of a measurement falls short of or exceeds the true measurement. It is not synonymous with mistake. It refers to the difference resulting from any source of inexactitude; it is involuntary. A mistake, on the other hand, may be voluntary or may arise from using a wrong method or making incorrect calculations.

The chief sources of statistical errors are three:

(1) **Errors of origin,** *e.g.,* errors resulting from a prejudiced information or faulty definition of units or due to erratic tendencies in the data;

(2) **Errors of inadequacy,** *e.g.,* errors due to inadequate sample data or incomplete information;

(3) **Errors of manipulation,** *e.g.,* unconscious error in measuring, weighing, counting or approximation.

Absolute & Relative Errors.

Errors may be measured as absolute or relative. An investigator is concerned with relative error more than with absolute one. **Absolute error is the difference between the true value and the estimate of a quantity,** while **relative error is the ratio of the absolute error to the estimate.** If the monthly average expenditure of students in a hostel is in reality Rs. 50, and we estimate it as Rs. 49, the absolute error is Rs. (50—49), *i.e.,* one rupee, while the relative error is Rs. $\frac{(50-49)}{49} = \frac{1}{49}$ = Rs. ·0204. The relative error is sometimes expressed as a **percentage error.** Thus, the relative error in the above case is 2·04 per cent. The error is positive since the true value exceeds the estimate. If, on the other hand, we estimated the monthly average expenditure as Rs. 51, the absolute error would be Rs. (50—51), *i.e., minus* one rupee, and the relative error:

$$\text{Rs. } \frac{50-51}{51} = \text{Rs. } \frac{-1}{51} = \text{Rs. } -\cdot 0196.$$

The percentage error would be

$$\frac{-1}{51} \times 100 = -1{\cdot}96\%.$$

The error is negative since the true value is less than the estimate.

In algebraic notation, let u represent the measurement of a quantity whose true value is u^1, and e stand for relative error of the estimate, then

$$e = \frac{u^1 - u}{u}$$

and, if ue stands for absolute error,

$$ue = u^1 - u$$

The error would be positive or negative according as u^1 is greater or smaller than u.

Biassed and Unbiassed Errors.

Errors may also be classified as biassed and unbiassed. **Biassed errors** result from a bias or prejudice on the part of the informant, enumerator or measuring instrument. These errors, therefore, lie in the same direction and are *cumulative* in character. That is, the greater the number of observations, the greater would be their absolute error. For example, if wall-paper were measured with a foot-scale half an inch short, the greater the number of feet measured the greater would be the absolute error.

Unbiassed errors are those which arise in the natural course or due to chance, that is without any bias or prejudice. They are subject to the law of statistical regularity, so that the excess in one direction is almost balanced by the defect in the other. Unbiassed errors, therefore, are *compensating*. The larger the number of items, the smaller will be their absolute error. If the foot-scale in our example be correct, error in a measurement in one direction shall be compensated by error in another measurement in the opposite direction, so that the greater the number of feet measured, the smaller will be the difference between the actual length and the length measured by the scale.

If some investigators carry on investigation into the economic condition of agriculturists in, say, the U. P. in India predetermined to prove that their income is high, they may examine only the well-to-do and debt-free cultivators and taking the incomes of those living near the grain markets, having their

own pack animals for transporting grain, and marketing their produce themselves, produce a high average income for each locality. But, if they are not prejudiced by a pre-determined conclusion, that is, if the inquiry is impartial, the investigators are as likely to make a low estimate in one locality as to make a high one in another. In the former case, the errors are biassed, all being in the same direction and pushing up the average. In the latter case, the errors are unbiassed, positive ones neutralizing the negative ones and reducing the resulting errors to a small figure. The following illustration would clear the point:—

Average monthly income in	True Estimate	Biassed Estimate		Unbiassed Estimate
	Rs.	Rs.	Rs.	Rs.
Locality A. ..	46·8	47	46	47
Locality B. ..	47·4	48	47	47
Locality C. ..	49·6	50	49	50
Locality D. ..	48·2	49	48	48
Averages ..	48·0	48·5	47·5	48·0
Relative errors ..	..	−1·03%	1·05%	Nil

From the above table it is clear that the errors of biassed estimate are cumulative, while those of the unbiassed one are compensating or counterbalancing. Another illustration of biassed and unbiassed errors is provided by the age-returns in the Indian Census. The fact that women generally under-estimate their ages causes a biassed error in the average age of women, while the tendency on the part of people to return their ages at the nearest round number causes an unbiassed error and, on the whole, does not affect the average age of the population materially. Biassed errors seriously affect the accuracy of the figures and should, therefore, be avoided. To eliminate the effect of unbiassed errors a large number of observations should be taken. So, if the errors cancel each

other even a considerable degree of inaccuracy may be allowed while editing the data, but if they are cumulative they shall have to be avoided.

Possible Error.

Possible error of an estimate represents the upper and the lower limits within which the actual error lies. If a quantity is rounded off to the nearest 100, the upper limit of error is $+50$ and the lower -50. The possible error is therefore expressed as ± 50. The figure 1449 or even 1450 can be expressed as 1400 if approximated to the nearest 100; and, similarly, the figure 1351 or even 1350 can be expressed as 1400 if rounded off to the nearest 100. If, therefore, we are given only the round figure 1400, and not the actual figure of which it is an approximation, we may express 1400 as $1400+50$, meaning thereby that the actual value of the estimate lies between 1450 and 1350, or that its possible error lies between $+50$ and -50.

Editing Secondary Data.

Even when all the precautions in using secondary data (outlined in the previous Chapter) have been taken it is necessary for one to exercise his common sense in scrutinising them before subjecting them to analysis. If the data are found to be lacking in consistency or homogeneity; if a time series in respect of a country does not cover the same geographical area from year to year, say, because of changes in its political boundaries; if the methods of collecting or computing certain data are known to be defective; if some coefficients are available but they had been computed without the numerators and denominators being homogeneous; if the available data are incomplete—in all such cases the data should be used only with great caution and care. Generally, all data are subject to certain limitations or qualifications which should be kept in mind by the editor and utilized by him for correcting the available statistics. Suppose the estimates of production of rice in West Bengal are based on the results of random sampling surveys from 1944-45 onwards and on the *Chaukidari* reports for previous years, some correction factor will have to be applied to previous years' data so as to make them comparable

with post-1944-45 data. Thus one has to constantly draw on one's knowledge about the nature, scope and limitations of secondary data in editing them. As Bowley observes: 'In collection and tabulation commonsense is the chief requisite and experience the chief teacher.'

EXERCISES

(1) What precautions should be taken in approximating large figures?

Approximate the following figures, expressing them (i) correct to nearest 1000 and (ii) correct to next 1000 over, and show the biassed and unbiassed errors in approximation:—

485,399; 410,902; 415,500; 290,492; 365,432; 399,491; 450,256; 462,300; 300,099; 295,591.

(2) Mention the kinds of errors likely to occur in the collection and interpretation of statistical data. What precautions would you take to avoid or minimise them?

(M.A., Alld., 1950).

(3) Mention the advantages of approximation in Statistics. What degree of accuracy is generally required in our statistical investigations?

(M.Com., Rajputana, 1951).

(4) What precautions should be taken in making use of published statistics for further investigation.

(B.Com., Agra, 1939).

(5) (*a*) Discuss the main sources of errors in statistics and their effects.

(*b*) State the important methods of approximation and their utility in statistics.

(B.Com., Agra, 1940).

(6) 'Let us have quantity as well as quality in statistical data; but if there be a choice between them, the latter is more important and essential than the former'—Explain.

Is the above a good maxim for one editing the collected data?

(7) In framing statistical estimates we are not so definite as The Modern Traveller who

"....knew the weather to a T
The longitude to a degree
The latitude exactly."

Examine the bearing of the above on the degree of accuracy desired in statistical estimates as distinguished from the estimates of the more exact sciences.

(M.A., Punjab, 1952).

(8) Write a note on the necessity of editing primary and secondary data before analyzing them.

(9) The statistician, who desires to safeguard his analysis and results from imperfections entering at the very start, should rest his choice among sources upon a test of reliability rather than upon accessibility and convenience.

Expand this statement so as to bring out clearly the way in which sources should be used.

(M.Com., Luck., 1943).

(10) Explain clearly, with illustrations, how a statistical error differs from a mistake. What classes of errors are known to you? How would you measure them?

(B.Com., Alld., 1949).

(11) Of the biassed errors the statistician should have none; but of the unbiassed ones, the more the merrier notwithstanding that they are also errors. Elucidate.

(B.Com., Alld., 1947).

CHAPTER VII

CLASSIFICATION AND TABULATION

CLASSIFICATION

Statistical data, collected in the course of an inquiry, concern a number of units of one kind or another which together form the group relating to the inquiry. A statistical group consists to a large number of things or individuals, having something in common, but differing from one another in respect of some measurable characteristics. For example, students belonging to the same college may differ from one another in regard to their age, civil condition or height. But, together they constitute a group. A statistical group is very large so that no one can appreciate at a glance, or even after a careful study, the information relating to its units. A reading of a thousand or more schedules returned by the students of a college respecting their age, weight, height, etc., cannot enable the reader to get a proper idea of the details mentioned. Some process of condensation must be devised for the purpose. This process yields statistical tables. But before tables can be prepared the different units must be grouped together into classes so that the like will go with the like and the unlike with the unlike. Details would necessarily be lost, since the individual units would be merged in a class. For instance, all those students who return themselves as 'married' shall be placed in one class, the 'unmarried' in another and the 'widowed' in the third. From the table that shall then be prepared none shall be able to identify himself or herself, since he or she shall be merely one in a class composed of those similar to him or her in respect of civil condition. This process is called Classification.

"Classification is the process of arranging things (either actually or notionally) in groups or classes according to their resemblances and affinities, and gives expression to the unity of attributes that may subsist amongst a diversity of individuals".[1]

[1] L. R. Connor, *Statistics in Theory and Practice*, 1926 ed., p. 18.

The objects of classification are many. It clearly shows points of similarity and dissimilarity. It helps one to form a mental picture of the objects he can see or conceive. By condensing the details it saves one from mental strain. It affords an appreciation of the information that would otherwise have been left out as perplexing or unimportant. It prepares the ground for enabling comparisons and analysis. It institutes a logical and orderly arrangement of things.

Importance of classification in statistics cannot be overemphasized, and yet it is something for which no very precise rules can be laid down. Skill and patience are, no doubt, indispensable; but as in collection of data so in their classification, experience alone will convince one of the requisite care that should be taken to avoid blunders and save time. It may be noted that an *ideal* classification should possess the merits of being unambiguous, stable and flexible. It should not leave room for doubt; it should be stable enough to render comparisons easy; and, it should be so flexible as to incorporate new ideas as they materialize in future.

Classification is determined by the characteristics possessed by the individual units of a group. These characteristics are of two kinds: descriptive and numerical. Descriptive characteristics comprize attributes or qualities possessed by objects or individuals, such qualities not being quantitatively measurable. Characteristics like sex, civil condition, caste, religion and infirmity are descriptive. Numerical characteristics are so called because they are susceptible of quantitative measurement. Age, height, income, weight are numerical characteristics. Classification of given data by descriptive characteristics is generally called classification according to attributes, while that by numerical characteristics is commonly known as classification according to class-intervals.

Classification according to Attributes.

Descriptive characteristics can be classified by means of some natural or physical lines of demarcation. Natural or physical differences determine the classes into which units

should be placed. It is easy in these cases to separate the similar from the dissimilar characteristics. For instance, population of India may be classed into male and female, literate and illiterate, blind and not-blind. Thus, when one attribute is noticed two distinct classes are formed. These two classes are exclusive of each other. If the members of one class possess the common quality of being males, those of the other are devoid of it. A classification of this type, where each class is divided into two sub-classes only, is called **Simple Classification** or **Classification by Dichotomy.**

Where more than one attribute is studied several classes may result. For instance, the population of India may not only be classed into males and females, but males and females may be further sub-divided into literate and illiterate such as male literate and female literate, male illiterate and female illiterate. Classification may be carried still further, for instance, according to occupation. A male literate may be a teacher, a female literate a stenographer; a male illiterate may be a peon, a female illiterate a maid-servant. Further classification according to religion or caste is yet possible. Numerous classes may thus be formed. A classification of this type, where each class is divided into more than two sub-classes, is called **Manifold Classification.**

The following brief classification of languages of India affords a good example of manifold classification:

A. *Languages of India and Burma.*
   (i) Austric, (ii) Tibeto-Chinese, (iii) Dravidian, (iv) Indo-European, (v) Unclassed.

B. *Languages of other Asiatic Countries and Africa.*
   (i) Indo-European, (ii) Tibeto-Chinese, (iii) Semitic, (iv) Hamitic, (v) others.

C. *Languages of Europe.*
   (i) Indo-European, (ii) others.

It is necessary to state that in classification according to attributes the boundary line between different classes, though artificially set, is definitely made before the work of classification begins. For instance, the decision as to who would be recorded as literate and who as illiterate is made before starting the actual classification of population according to literacy.

Classification according to Class-Intervals.

Numerical characteristics can also be classified by assigning arbitrary limits. The ages of persons, for instance, vary; so do their heights or weights. But, the entire range of ages, heights or weights, from the lowest to the highest, can be broken up by drawing arbitrary boundary lines, and those units which are nearly alike in respect of a particular character put together in one class. Thus if the ages of a given group of people vary from 25 to 44 years and it is desired to divide the group into four classes, the boundary lines would preferably be fixed on numbers 25, 30, 35, 40 and 45 years. These boundary lines are known as the **class-limits**, the group constituted by two limits as the **class-interval**, the distance between two limits of a class-interval as its **magnitude**, and the number of observations falling within a particular class-interval as its **frequency**. In our example, we group together those whose ages are 25 years and more but less than 30 years and place their number against the class-interval 25-30 years, group those who are 30 years and more but less than 35 years and place their number against the class-interval 30-35 years, and so on. Evidently, the magnitude of our class-interval is 5 years. Unit is the most common magnitude. So far as possible, the magnitude of all class-intervals should be uniform, so that the labour of calculating different statistical constants may be minimised. Even if a particular class-interval contains no frequency, the class-interval must be entered in its proper place; otherwise errors might be made in plotting the results. The limits of the class-intervals should preferably be so fixed that the mid-point of each falls on an even unit and not on a fraction. One might ask—'How many groups should there be ?' In answer it may be said that the number of class-intervals is dependent on the nature of the inquiry. In general a number of groups in the neighbourhood of 20 is the most satisfactory, provided the number of observations is reasonably large. Thus classification according to class-intervals is obtained when a numerical characteristic is considered and each group is subdivided into a number of classes or groups rather arbitrarily.

The above class-intervals, *viz.*, 25-30 years, 30-35 years, etc., were expressed according to the **exclusive method**, that is,

the upper limit of the one class-interval was the lower limit of the succeeding class-interval. An item 29·99 years would fall in the class-interval 25-30 years, while an item exactly 30 years would be taken to the second class-interval (30-35 years). This difficulty can be won over by classifying the class-interval as 25 and under 30 years,' '30 and under 35 years,' and so on.

Class-intervals are also expressed by the **inclusive method.** The above class-intervals (arranged according to the exclusive method) would be expressed as 25-29 years, 30-34 years, etc., according to the inclusive method. In this case the upper limit of a particular class-interval is also included in the class-interval itself. The first class would include all items between 24·5 and 29·5 years. To be still more unambiguous, the class-intervals may be expressed as 25-29·9 years, 30-34·9 years. But the inclusive method is not in general use since in adopting it the idea of continuity in the limits of class-intervals is lost.

Statistical Series.

If the quantities or values of some aggregate are measured, counted or weighted, or numbers in some group or class are counted, and they are placed one after another, the result is a statistical series. Briefly, a **statistical series** may be defined as things or their attributes arranged according to some logical and systematic order.

Time, Spatial and Condition Series.

There are three bases of classification of data: time, space and condition. Accordingly there are three types of statistical series: (1) **historical** or time (2) **spatial** and (3) **condition.** In the first, facts are arranged with respect to time; for instance, index numbers of wholesale prices of wheat in India over a period of time detailed in chronological order. In the second, the controlling factor in presentation is place: variations are noted geographically. Production of wheat in India for a given date arranged according to different States would constitute a spatial series. In the third, variations in size and amount of things or their attributes are shown. The different measurements of natural phenomena are usually distributed about a norm. If heights, weights, or ages of students, or

lengths of a number of leaves chosen at random are measured, the different measurements, when arranged in logical order, shall constitute a condition series, and it will be noticed that though the different measurements would vary, a most common, predominant weight, height, age or length of leaves would be found. We shall see later (in the Chapter on Statistical Averages) that this 'most common' length is called the 'mode' of the series. Condition series takes the form of what is called 'frequency table'.

Continuous and Discrete Series.

Series may be continuous or discrete. When the items of a series are not capable of being determined with mathematical accuracy, but are always measured by approximation and can only be placed within certain limits, the resultant record is a **continuous series.** On the contrary, where the items are exactly measurable and their record shows definite breaks between one value and the other succeeding it, the resultant record is a **discrete** or broken or **discontinuous series.** For illustration see Table 5.

Measurements of weights, magnitude and volume constitute continuous series, for apparently there is no limit to the sub-division of maunds, miles and gallons. The numbers of labourers in factories, of spots in dice-throws or of pages in books form discrete series, since they must give integral numbers and are incapable of sub-division.

TABULATION

After the data have been classified they may be tabulated, that is, put into tabular form. **Tabulation stands for the systematic and scientific presentation of quantitative data in such a form as to elucidate the problem under consideration.** Its function is to arrange in an orderly manner the answers to those questions with which the inquiry is concerned. Tables are intended to summarize the information obtained in course of an investigation.

Rules and Precautions for Tabulation.

Some precautions in drawing up tables are necessary. The

given data may be grouped in one table or several tables. A single table shall, no doubt, bring the entire data into proximity; but if it is too large, it shall confuse the eye and lead to great difficulty in following the columns and rows at a glance. To do away with such inconvenience it may be broken up into several separate tables. Further, several comparisons of different nature should not be jumbled up in one table. Each table should be a unit. Usually there should be separate tables for different distinct purposes. Again there should be few main divisions with several sub-headings under each. If the number of headings is very large, the main facts to be compared may not be adequately emphasized. Of course, the exact number of divisions and sub-headings shall be determined by the data in hand. Each table should be so complete in itself that it may not be made more intelligible by re-drafting. The table should suit the size of the paper on which it is drawn. So, the width of each column and row should be properly calculated and headings correctly arranged before the permanent table is ruled or figures are entered in it. Totals, averages, percentages and the numbers that are to be compared should be placed close together, and, if possible, they should be placed in the same vertical column rather than the same horizontal row. Columns that are to be compared should be placed adjacent to one another. The rulings in tables should be such that principal groups are separated by thick or multiple-ruled lines. Unimportant data may be grouped together and placed in 'miscellaneous' group. Items which are in any way different from the rest of the items—*e.g.*, estimated figures, revised figures—should be marked with an asterisk or number, and an explanatory note given beneath the table. The table should be given a suitable title. This title and the titles of sub-headings should be self-explanatory, so that no reference to the text or footnotes may have to be made. The title should neither be too small nor ambiguous. The column-heading should indicate the unit used, as 'height in inches,' 'price in rupees', 'weight in tons'. Large digits may be approximated and mentioned in thousands, lakhs or millions. A table may show absolute figures, increase or decrease from past years' figures, percentages, etc., according to the nature of comparison to be made; any qualifications to which the data may be subject

should therefore be indicated in suitable footnotes. For each column the source or the manner of arriving at the figures given in it should preferably be stated below the table. All items should be carefully checked before entering in the table or totalling them up. Arrangement should also be made in the table for cross-checking. There should be no over-writing, otherwise the neatness of the table would be lost. A note pointing out the principal conclusions and their limitations should accompany the table.

Different types of Tabulation.

In very general terms tabulation may be distinguished as simple and complex. A **simple table** contains data respecting one characteristic only, information relating to other characteristics being left out. Table 1 is a case of simple tabulation. A more **complex table** may contain figures relating to several characteristics. Tables 2, 3 and 4 represent this type.

Tabulation is also classified as Single, Double, Treble and Manifold. A **single tabulation** is one that answers one or more groups of independent questions. The following table gives the frequency distribution of marks obtained out of a maximum of 50 by the students of a class in their test in Economics:

Table 1. *Frequency Distribution of Marks in Economics.*

Marks-group	Number of students (frequency)
0-5	4
5-10	6
10-15	10
15-20	16
20-25	12
25-30	8
30-35	4

The table is capable of furnishing the first approximation to the answer to an inquiry into the 'ordinary marks' obtained by the students in the test. It tells, for instance, that the

number of students getting marks between 15 and 20 is the highest as compared with the number in any other group. A still simpler table will be one showing yearly variations of some characteristic, say, progress of cotton mill industry in India or of trading profits of a company.

Double tabulation shows the sub-division of a total according to two categories, and is capable of answering two mutually dependent questions. Table 2 is an illustration of this type of tabulation. It shows the distribution of Route-mileage of Class I (Government managed) Railways in India in the middle of April 1952 into (i) Railway Zones and (ii) Gauges.

Table 2. *Route Mileage of Class I Railways in India.*

Railways	Route Mileage Open (as on 14-4-'52)			
	Broad Gauge	Metre Gauge	Narrow Gauge	Total
Central ..	4,091	773	564	5,428
Eastern ..	4,725	..	942	5,667
North Eastern ..	2	4,713	52	4,767
Northern ..	3,882	1,998	128	6,008
Southern ..	1,755	4,160	102	6,017
Western ..	1,266	3,402	793	5,461
Total Class I ..	15,721	15,046	2,581	33,348

Treble tabulation sub-divides a total into three distinct categories and answers three mutually dependent questions. Table 3 illustrates treble tabulation. It shows the distribution of India's population in 1951 into livelihood categories, by the two sexes, for the groups of States as mentioned in the Constitution.

Table 3. *Livelihood Categories by Groups of States—1951 Census.*

State Groups (as per Constitution of India)	Agricultural			Non-Agricultural			Total		
	Male	Female	Total	Male	Female	Total	Male	Female	Total
Part A ..	99·7	97·6	197·3	43·4	38·0	81·4	143·1	135·6	278·7
Part B* ..	23·1	22·1	45·2	11·7	11·0	22·7	34·8	33·1	67·9
Part C ..	3·3	3·2	6·5	1·9	1·6	3·5	5·2	4·8	10·0
Part D ..	0	0	0	0·2	0·1	0·3	0·2	0·1	0·3
Total India*	126·1	122·9	249·0	57·2	50·7	107·9	183·3	173·6	356·9

* Excluding Jammu and Kashmir.

Manifold tabulation is one that divides a total into several categories, generally more than three. The following specimen blank table drawn to show the distribution of population in the various States of India according to age, sex, literacy and caste illustrates manifold tabulation.

Table 4. *Distribution by age, sex, literacy and caste in India.*

State	Caste	Age-Group	Male			Female			Total		
Assam	Kayasth	0-25									
		25-50									
		50-75									
		Over 75									
		Total									
	Brahmin	0-25									
		25-50									
		50-75									
		Over 75									
		Total									
	All Castes	0-25									
		25-50									
		50-75									
		Over 75									
		Total									
Bihar	Kayasth	0-25									
		25-50									
		50-75									
		Over 75									
		Total									

EXERCISES

(1) Define Classification and Tabulation and show their importance in statistical studies.

(2) "In collection and tabulation common sense is the chief requisite and experience the chief teacher"—Bowley.

What precautions, in your opinion, are necessary to avoid statistical errors in the collection and computation of economic data ?

(M.A., Agra, 1940).

(3) "Classification is the process of arranging things (either actually or notionally) in groups or classes according to their resemblances and affinities, and gives expression to the unity of attributes that may subsist amongst a diversity of individuals."—Connor.

Elucidate the above statement.

(B.Com., Alld., 1947).

(4) (*a*) Draw up a blank table with suitable headings, spacings, double lines, etc., in which could be shown the number and tonnage of ships entered and cleared at ports in India, for 10 years, distinguishing steam and sailing vessels, and also those with cargoes from those in ballast.

(*b*) What do you mean by a "statistical unit of measurement ?" Give a suitable illustration.

(B.Com., Hons., Andhra, 1942).

(5) Re-arrange the following blank table with a view to make it more intelligible.

Sex	Brahmin		Rajput		Kayastha		Harijan	
	Literate	Illiterate	Literate	Illiterate	Literate	Illiterate	Literate	Illiterate
Male								
Female								

(B.Com., Alld., 1940).

(6) Draw up in detail, with proper attention to spacing, double lines, etc. and showing all sub-totals, a blank table in

which could be entered the numbers occupied in six industries at two dates distinguishing males from females, and among the latter single, married and widowed.

(M.A., Alld., 1940).

(7) Prepare a specimen form in blank, with suitable heading and spacing, for use in collection of data on *one* of the following:—

(*a*) Survey of trades in your district.
(*b*) Standard of living of middle class families in a small town.
(*c*) Expenses of students in a University.

(Dip. in Econ., Madras, 1931).

(8) Explain how you would tabulate statistics of death from principal diseases by sexes in different provinces of India for a period of five years.

(B.Com., Cal., 1937).

(9) Prepare a table with a proper title, divisions and sub-divisions to represent the following heads of information:—

(*a*) Exports of cotton piece-goods from India.
(*b*) To Burma, China, Java, Iran, Iraq.
(*c*) Amount of piece-goods to each country.
(*d*) Value of piece-goods to each country.
(*e*) From 1939-40 to 1945-46, year by year.
(*f*) Total amount exported each year.
(*g*) Total value of exports each year.

(B.Com., Alld., 1946).

(10) Discuss the function and importance of tabulation in a scheme of investigation.

Prepare blank tables, showing the distribution of the students of a University according to age, class, and residence, for arranging (a) physical training, and (b) seminar classes.

(11) What precautions would you take in tabulating your data ? Prepare a blank table to show the distribution of population according to sex and four religions, in five age-groups, in seven important cities of U. P.

(B.Com., Agra, 1937).

(12) Draw up two independent blank tables, giving rows, columns, and totals in each case, summarizing the details about

the members of a number of families, distinguishing males from females, earners from dependents, and adults from children.

(M.A., Cal., 1935).

(13) Tables are a means of recording in permanent form the analysis that is made through classification and of placing in juxtaposition things that are similar and should be compared. (Secrist)—Explain.

(14) The State of Rajasthan had only one city with population of a lakh and over in 1921 and 1931. It was Jaipur. In 1941 two more cities—Jodhpur and Bikaner—were added. 1951 was no further additions. It is significant that in 1941 the two new cities had identical population, which was exactly the same as that of Jaipur twenty years ago. The total population of the three cities in 1941 was 430 thousand, of which about 40% (the true figure being 176,000) was accounted for by Jaipur alone, the remainder being shared equally by the other two cities. In 1941, while Jaipur and Jodhpur added respectively 26,000 and 32,000 to their 1931 numbers, the addition in case of Bikaner was much larger, being 15 thousand more than that in Jaipur. In 1921, the combined population of Jaipur and Jodhpur was just two lakhs, and Bikaner had four thousand souls less than Jodhpur. The effect of events like Partition and integration of states on the size of these cities was reflected in the 1951 census figures. In that year, as compared with 1941, Bikaner registered a decline of 10,000 persons, Jodhpur an increase of 47,000 souls, and Jaipur an increase which was only two thousand less than the total population of Bikaner in 1951 or which was five times the increase Jaipur registered between 1921 and 1931. It will be seen that during the thirty years intervening 1921 and 1951, the growth of population in Bikaner was only 48,000, while that in Jodhpur was 101,000 and that in Jaipur higher still—164 thousand. It will also be seen that while Jaipur continued to enjoy the distinction of being Rajasthan's city No. 1, Bikaner, which occupied the 2nd place in 1941 along with Jodhpur, gave that place to Jodhpur in 1951 and got relegated to the third.

From the above account prepare a table showing the population of the three cities for the different census years and

calculate the rate of increase or decrease in their population in 1951 over 1941.

(M.Com., Rajasthan, 1952).

(15) Following are the weekly earnings of labourers in a factory:—

Earnings Rs. a. p.	No. of labourers
4 6 0	25
10 10 6	6
5 9 3	35
5 1 0	42
6 11 9	30
7 12 3	21
9 9 0	17
8 2 0	15
10 0 0	8
5 0 0	52

Tabulate the above data in classes of Rs. 4-5, Rs. 5-6, etc., and give a suitable heading to the table.

(16) How will you proceed to classify the observations made, and what points will you take into consideration in tabulating them ? Mention the kinds of tables generally used.

(B. Com., Agra, 1941).

CHAPTER VIII

SIMPLE DERIVATIVES

After the collected data have been edited, classified and tabulated, the resulting table, though compressing an unwieldy mass of figures to a large degree, shall not be easily grasped or compared with other tables because it would contain a number of entries. Some method of concisely describing the data has, therefore, to be devised. The most simple method is that of computing certain derivatives from the data. **A statistical derivative is a quantity resulting from a combination of two or more original figures.** Therefore, it should be remembered that statistical derivatives do not arise from simple measurement or counting, but always from computation.

Derivatives are both simple and complex. We shall later see how statistical averages of the first order[1] are derived from the original data. These averages are **complex derivatives**, and so are those of the second order—the measures of dispersion[2]. **Simple derivatives** consist of relative numbers. Two relationships are distinctly marked: Subordinate and Co-ordinate. Accordingly, there are subordinate and co-ordinate derivatives.

Subordinate Derivatives.

Subordinate derivatives are those which show the relative size of parts to a whole. They are generally expressed as **proportions** or **percentages**, e.g., the proportion or percentage of area under food-crops and under non-food crops to total area cultivated. Here the total is divided into two categories.

Co-ordinate Derivatives.

Co-ordinate derivatives are those which show the relative size of pairs of inter-related co-ordinate masses. These include several varieties:

[1] See Chapter IX, on *Statistical Averages.*
[2] See Chapter X, on *Dispersion and Skewness.*

1. **The Simple Difference** between two quantities of like kind, *e.g.*, comparing current year's production of sugar in India with last year's.

2. **The Percentage Difference**, the difference being expressed as a percentage upon some quantity taken as standard, *e.g.*, current year's production is so much per cent. higher or lower than that of the last year's.

3. **The Ratio.**—It is another way of expressing the percentage difference. Instead of saying what we did in case of percentage difference, we may say that production of sugar in the current year has increased from that of the last year in the ratio of 100:120, or 50:60, or 5:6, all ratios being identical ways of expression.

4. **The Rate.**—Generally speaking, when the two quantities to be compared are of the *same* kind we use the term ratio, *e.g.*, ratio of boys to girls in a university. Here both the quantities are the same, *viz.*, students of the same institution. But when the numerator and denominator are of *different* kinds we speak of rates, *e.g.*, sickness rate, marriage rate, mortality rate. Here the comparison is between quantities of different kinds.

Further, a rate is usually standardized in regard to the denominator. One mass is divided by the other related mass and the quotient is multiplied by 100 or 1,000; and, we speak of rate per cent., rate per *mille*.

But the distinction between rates and ratios is not rigid. We speak of birth rate, *i.e.*, number of births per 1,000 population. We may equally correctly speak of the birth ratio, *i.e.*, ratio of the number of births to the number living.

Rate per unit is called **statistical co-efficient.** If the birth rate is 30 per 1,000, the co-efficient is .03. The characteristic of this co-efficient is that if it is used to multiply a total, (*e.g.*, population) an allied number (number of births) would be obtained.

Purpose of Computing Statistical Derivatives.

Simple derivatives are computed to compare statistical groups. In the computation of rate per cent. or rate per *mille* observations are reduced to a common denominator and com-

parison is thereby facilitated. If in University A 900 candidates were successful out of 1200 who appeared, and in University B 980 passed out of 1400, a camparison of the absolute number of successful candidates—900 and 980—without considering the number of those who appeared at the examination, would lead one to declare the result of the B University as better than that of the A. But, when the two results are reduced to a common denominator, say, expressed in percentages, this impression will be reversed. The percentage of successful candidates in A is,

$$\frac{100 \times \text{Number of successful candidates}}{\text{Number appeared}} = \frac{100 \times 900}{1200} = 75.$$

Similarly, the percentage in B is,

$$\frac{100 \times \text{Number of successful candidates}}{\text{Number appeared}} = \frac{100 \times 980}{1400} = 70.$$

Both the results have now been reduced to a common denominator, 100. It is evident that the percentage of success in A University is higher than that in B University. Therefore, A's result is better than B's. The usefulness of relative numbers for purposes of comparison is thus clear.

But relative numbers can also be used for another purpose, *viz.*, computing the size of an unknown mass from a known one. The known mass, *i.e.*, the relative number, may be an actual figure or an estimated one. Relative numbers are often estimated when there are no sufficient data for their computation. Such estimates can be employed to know the size of the mass to which they relate. Statisticians have often used them to obtain the population figures of past times. If we know, historically, the number of artisans or beggars of a city or country, we may make an estimate of the percentage of the entire population that the artisans or beggars probably formed, on an average, at the time under consideration, and hence compute the total population for that time. The estimation of the artisans or the beggars is made possible by the law of statistical regularity: the ratio between definite statistical masses is often fairly constant. This ratio can be easily estimated to be within certain limits. This holds good, for example, for the relationship between population and births and deaths. Every estimate, however, must be regarded as

simply an approximate value. It may or may not be accurate. Therefore, the size of an unknown mass computed from an estimated relative number must also be regarded as merely approximate.

But, where actual percentages are known the mass or population to which they relate can be ascertained to a considerable degree of precision. For, given that the number of successful candidates at a certain examination was 900, and this constituted 75% of the total, it is easy to see that the total number of candidates who appeared at the examination was 1200.

Derivative Series.

A set of relative numbers or simple derivatives of the same kind spread, say, over a period of time would constitute a derivative series. The special feature of a derivative series is that it eliminates the factor or factors obstructing effective comparison: all figures are related to a common denominator and comparison is facilitated. A number of figures representing burden of income-tax per head of population over a period of time forms a derivative series. The series would eliminate the main effects of demographic changes. Population on which the burden is computed may change, yet the burden per head of population for one year shall be comparable with similar burden for another year. The actual figures of population and amount of tax would not be so easy to compare from year to year because of their fluctuations.

The test of a derivative series lies in its stability which is determined by meaures of dispersion. These measures shall be discussed later (in Chapter X); but it will be useful to note here that higher the degree of stability, greater is the reliability of a derivative series. To attain stability some simple precautions should be kept in view in computing simple derivatives.

Precautions in computing and using Derivatives.

Both, the computation and the use of relative numbers, need caution. The rule should be to procure as much homogeneity in the data as possible. For example, general death-rate for a town or a country may be computed by multiplying the number of deaths by 1,000 and dividing the product by the

total population; but, this general death-rate, or crude death-rate as it is called, relates to heterogeneous mass, since deaths vary with age and sex compositions of the population. We shall later see (in Chapter IX) how this defect can be remedied.

In order that there may be no misunderstanding, the basis of calculation of the relative numbers should always be given. If we are told that the price of a commdity increased 10 per cent., decreased 15 per cent., increased 25 per cent., decreased 20 per cent., and then increased 15 per cent., over a period of time, it would be difficult to say what exactly the change over this period was. If the changes were based on the original price, it would be found that the change over the whole period was 15% on that price. If, however, the changes were based on the prices ruling at the time of each particular change, the change over the period would be found to be about 7.5% of the basic price. If the basis on which percentages were calculated were known, this variation in result would not occur.

Again, percentages, or other relative numbers, should be used only when the factors which are to be expressed in percentage form are themselves comparable. If a company whose issued and paid up capital was Rs. 150,000 earned profits at a uniform rate of Rs. 15,000 per year for five years, its percentage of profits to capital would be 10 for the period. Suppose the company increased its capital to Rs. 250,000 in the sixth year and the profits increased from 15,000 to Rs. 22,500 in that year, the percentage of profits to capital wuld be 9 only. Then, if a table showing only the percentage profits were prepared for the six years, it might lead one to conclude that profits had declined in the sixth year while, in fact, the actual profits had increased. In such cases it is advisable to show the amount of capital over the different years, the total profits earned from year to year and the rate of percentage profits. This would avoid all fallacy.

Lastly, percentages should not ordinarily be used when the number of items in one of the series to be compared is less than one hundred. Similarly, rates per thousand or rates per ten thousand should not be computed when the number of items is comparatively very small. Advertisements very often appear in the newspapers: "Join X school. This year's results 100 per cent." Another institution may have a percentage of only 92.

A prospective candidate may be led to think better of the first institution. But, if on an inquiry it were found that only 3 candidates appeared from X school and all got through, while 250 candidates appeared from the second institution of which 230 came out successful, opinion will have to be reversed, for it is always more difficult to get the same percentage result from a much larger number. Mathematically both the percentages are absolutely correct, but *statistically* the percentage relating to the result of X school is not significant. It is, therefore, again established that comparison through percentages alone is not sufficient unless the data on which they are based are homogeneous and capable of comparison.

To eliminate the chances of fallacious conclusions, which might result from the use of percentages when their bases of calculation are not specified, some statisticians emphatically recommend the use of ratios in place of percentages. Then we shall say that the price of the commodity increased in the ratio of 100.110 rather than that it increased 10%.

But ratios must also be used with caution, otherwise wrong inferences might be drawn. Suppose 800 candidates appeared at a certain examination of which 600 came out successful. The ratio of passes to failures is, therefore, 6:2. Further supposing college A *coached* 500 out of the 800 candidates, and of these 400 passed so that the ratio of the successful to the failed candidates is 4:1, whereas of the remaining 300 candidates 200 must have passed which gives the ratio of successes to failures 2:1. It might appear that college A achieved twice as good results as all those colleges taken together through which the remaining 300 candidates appeared. This conclusion is fallacious as it is not known whether all the 300 students were given adequate coaching or they simply *appeared* through certain colleges. If they were not properly coached they did not stand the same chances of success as those 500 candidates who were given due coaching. It is advisable, in such cases, to show the ratios of all the coaching institutions in order to ascertain which one of them was really the best.

The use of rates has become very common in population studies in modern times. The growth of population is determined by the mortality rates, fertility rates and the age composition of the population.

Mortality Rate.

Mortality rate means the number of deaths per thousand babies born. If out of 1,000 children born, 150 do not survive, the mortality rate is 150 per thousand. The survival rate is 850 per thousand.

Fertility Rate.

Fertility rate refers to the number of babies born per thousand women of the child-bearing age. If on an average 120 children are born per thousand women of 20-25 years of age, the fertility rate per thousad women of the given age-group will be 120.

Gross Reproduction Rate.

The gross reproduction rate gives the number of female babies born per thousand women of child-bearing age. Suppose the child-bearing age in women ranges between 15 to 50 years in a country and the following fertility rates in respect of the different age-groups of child-bearing women are available:—

Age-Group of child-bearing women.	*Fertility rate per 1,000 women.*
15-20	20
20-25	170
25-30	260
30-35	200
35-40	160
40-45	70
45-50	10
	890

The fertility rates given above are the averages for the five-year periods. Total fertility is, therefore, equal to 800 × 5 or 4,450. This is the total number of babies, both male and female, born per thousand women. Assuming that the ratio of female births to total births is 49% in the country concerned, it follows that out of 4,450 babies born, 2,136 were female. The gross reproduction rate is, therefore, 2,136 per thousand. Gross reproduction rate is generally expressed per unit. Therefore, in our example, the gross reproduction rate is 2.136. This signifies that for one present mother there will be 2.136 in the future; in other words, it shows how the population is reproducing itself.

Net Reproduction Rate.

In the above interpretation of gross reproduction rate, viz. for one present mother there will be 2·136 in the future, there are two assumptions: One is that none of the female babies born would die before reaching the age of 50 which is the upper limit of child-bearing age in our example, and the other is that fertility rate would remain unchanged. These assumptions may not come true. If allowance is made for mortality among the babies born, it is obvious that for one present mother there will not be 2·136 in the future. Similar will be the case if the fertility rate with which we started in our example did not apply to succeeding generation, that is, if fertility rate did not remain constant. If, therefore, the gross reproduction rate is corrected for these variations, the resulting rate is the net reproduction rate. It is clear that the net reproduction rate would be lower than the gross. If the net reproduction rate is 1, it shows that for one present mother there would be exactly 1 in the future, that is to say, the population of the country is stationary. If the rate works out to less than 1, it shows a declining population; and, if it is greater than 1, it means the population is growing.

Let us take an example of growing population:—

Age-group of child-bearing women.	No. of female children born to 1,000 women passing through each age-group.	No. of survivers out of each 1,000 female children.	No. of surviving women by whom present women replace themselves.
15-20	50	850	42·5
20-25	200	800	160·0
25-30	600	750	450·0
30-35	500	700	350·0
35-40	450	650	292·5
40-45	150	600	90·0
45-50	40	500	20·0
	1,990		1,405·0

In all, 1,990 female babies were born to 1,000 women giving a gross reproduction rate of 1·99. Assuming that the fertility rate remains constant, 50 female babies, when they attain the age of 15 to 20 years, will give birth to 50 female babies. But not all of the 50 female babies will survive. Assuming a survival rate of 850 out of 1,000 for the female babies born to women of the age-group 15 to 20, only 42·5 will survive. Thus, the total 1,990 declines to 1,405, after allowing for mortality. That is, the place of 1,000 women is taken not by 1,990 women, but by 1,405 women. The net reproduction rate is 1·405; for one present mother there will be 1·405 in future. The popula-this is thus growing.

Use of Simple Derivatives.

Ratios, rates per unit, per hundred, per *mille* are widely used and easily understood. Sex ratio, cost per unit of output, income *per capita,* percentage rate of interest, percentage of exports or imports to total trade, birth and marriage rates per thousand, mortality rate per ten thousand, burden of tax per head of population, net reproduction rate—these derivatives suggest the variety of field to which they are applied. They are very commonly used in business, social and administrative statistics.

EXERCISES

(1) Define a statistical derivative, and point out the usefulness of its computation in statistical studies.

(2) Clearly distinguish between Subordinate and Co-ordinate derivatives.

(3) What is a derivative series ? How does it differ from (*i*) series of individual observations and (*ii*) frequency distributions ?

(4) What rules and precautions will you observe in computing percentages and ratios ?

(5) Explain what you understand by (*a*) amount of tax per tax payer, (*b*) burden of tax per head of population, (*c*) net reproduction rate, (*d*) income *per capita,* (*e*) yield per acre, (*f*) cost per unit of output.

(6) Point out the ambiguity or mistake, if any, in the following statements:—

(*i*) The death-rate in the American navy during the Spanish-American war was nine per thousand while in the city of New York for the same period it was sixteen per thousand. It was safer, then, to be a sailor in the American navy than to live in New York City.

(M.A., East Pb., 1951).

(*ii*) 99% of the people who drink die before reaching 100 years of age. Therefore, drinking of wine is bad for longevity.

(M.A., Patna, 1943).

(*iii*) The rate of increase in the number of cows in India is greater than population. Therefore, the people of India are now getting more milk per head.

(M.A., Patna, 1943).

7. What statistical data are necessary for the calculation of the net reproduction rate ? What is the deficiency in the existing Indian data in this respect ?

(M.A., Alld., 1951).

8. What is net reproduction rate ? Explain with the help of an example the method of calculating it.

(M.A., East Punjab, 1950).

9. Below is given the fertility rate per 1,000 women by their age-groups for a certain country for 1936:—

Age-group	*Fertility rate per 1,000 women*
16-20	19
21-25	173
26-30	253
31-35	201
36-40	157
41-45	67
46-50	9

Assuming that the ratio of female babies to total births for the country and year concerned is 48·8 per cent, calculate the gross reproduction rate for that country and explain what this rate means ?

(M.A., East Punjab, 1951).

10. What are the various ways of the measurement of population growth ? In this connection discuss in detail the calculation of net reproduction rate.

(M.Com., Alld., 1952).

CHAPTER IX

STATISTICAL AVERAGES

Simple statistical derivatives, by themselves, are insufficient to give a summary description of the characteristics of a large group of variables, nor can they be used as types representing the group. They throw light only on the relative aspects of a series, and are not representative of the data. Therefore, some other method of precisely and concisely describing the series has to be devised. This is the method of statistical averages. Through it a representative or characteristic of the entire group is computed, which affords the central idea of the group and can be used in place of it.

Statistical Average and its Object.

An average is a single simple expression in which the net result of a complex group or large numbers is concentrated. Averaging is thus the process of condensation. The average brushes off the irregularities of a series, levels all differences of the individual items and presents complex and unwieldy data in a significant number. It thus gives a bird's-eye view of an aggregate, and can be substituted for individual items in further calculations regarding the series. It is a typical item representing the group and is, therefore, also called **type.** It is a representative number of the group, its gist. As such, it describes the group or throws light on its characteristics better than any other value in the group.

There are two main objects of computing a statistical average. One is **to give a concise picture of a large group,** to describe the series it represents. The other, which follows from the first, is **to afford a basis of comparison with other groups.** It follows that an average may be computed for its own sake, or as a means to another end which may be comparison, or measurement of dispersion[1] or of skewness[2] of the series. It is difficult to grasp the significance of the ages of the inhabitants of a country if we are given the age of every person in the country, but the average age of the people of the country is

[1,2] For Dispersion and Skewness see Chapter X.

something definite and intelligible. Similarly, two series each containing ages of different people in two countries, even if tabulated into a few frequency classes, will not afford a comparison between the ages of the people in the two countries. If, however, some sort of average of the two series is computed, average ages in the two countries shall be comparable at a glance. And, comparison of the average ages shall usually be equivalent to comparison between the two age-series.

Law of Averages.

The principle underlying the use of averages is that over a long time, distance or attribute possessed by a large number of cases, abnormalities in one direction are generally offset by those in the other; and the average, by offsetting these abnormalities and by being present in the data, gives a single, simple, summarised expression of the whole data. This is the law of averages. In terms of this law, an average of a group of variables is the value around which the variables centre, the value which balances the deviations of the variables from it. The average of 4, 5, and 6 is 5; deviations below and above 5 would balance (4 − 5 balances 6 − 5).

Homogeneity of Data.

It is necessary that the data from which averages are computed are as largely homogeneous as possible. For, if the different items are not alike in relevant aspects there is no sense in grouping them together, and consequently, no justification for computing their averages. The comparison of only those averages yields reliable conclusions which refer to homogenous masses. The significance of average lies in the fact that it exhibits the result of the activity of complex causes in one characteristic figure. The average wage of a given group of workers, for instance, gives a measure of the factors determining wages in that group. It is important that this average should refer to as unified complex of causes as possible, since then alone will it be reliable for purposes of comparison or as a type. If the wages in two factories are determined by quite different causes, the average wages for all the workers in the two factories shall not yield a trustworthy comparison. Hence the importance of homogeneous series.

Homogeneity can be attained by (1) eliminating the unlike from the like items and (2) dividing the like items into groups as nearly homogeneous as possible. In our example, it will be necessary to disregard the cases in which workers for personal reasons do not receive any wages at all. Among those getting wages, workers may be distinguished whose wages are influenced by different and independent causes. They will have to be divided into more homogeneous parts, so that the averages may refer to unified complex of causes, and thus be reliable for comparison between two periods or two places. Our wage-earners may consist of males and females, the latter getting lower wages. Therefore, male workers will have to be separated from female workers, and separate averages computed. There is still a possibility of these two groups being sub-divided into more homogeneous parts, *e.g.*, skilled male workers and unskilled ones.

The extent to which homogeneity should be attained is determined by the purpose for which averages are required. We hear of the average tax per tax-payer as also of average tax per head of population. Both have different purposes: the first is a measure of the burden on the tax-payers only, the second of that on the whole population.

Kinds of Average.

The various kinds of averages may be grouped as under:

I. AVERAGES OF POSITION:
   (1) Mode, and
   (2) Median.

II. MATHEMATICAL AVERAGES:
   (1) Arithmetic Average, or Mean,
   (2) Geometric Mean,
   (3) Harmonic Mean, and
   (4) Quadratic Mean.

III. OTHER OR BUSINESS AVERAGES:
   (1) Moving Average, and
   (2) Progressive Average.

Averages of position represent rather an estimate of central tendency than a strict mathematical concept which the

Mathematical Averages do. Of the above averages, Harmonic and Quadratic means and Progressive and Moving averages are not in common use.

MODE

Mode is the value of that item in a variable which occurs most frequently or is repeated the greatest number of times. It lies at the position of greatest density. It is the typical measurement or most fashionable point. It is the usual, and not casual, size of item in a series. When we speak of the average student, the average wage, the average rent, etc., we generally imply the modal student, the modal wage, the modal rent. If we say that modal marks obtained by students in a class are 40, we mean that 40 are the predominant marks, *i.e.*, the largest number of students secured 40 marks. As high as 70 marks and as low as 15 marks in this particular case are exceptions; they are much less frequented; they are non-modal.

Since mode is the most frequent size, it appears it is easy to locate it. Really it is so, if there is a single well-defined mode. Then the size of item, or the group, containing the maximum frequency, can be easily located in a frequency table. But, it is not improbable that there may be numerous irregularities in the table, so that the position of the mode would become indefinite and modal size would not be easy to locate. In such cases frequencies are adjusted by the process of grouping, *i.e.*, by widening the groups into which the frequencies fall until a mdoal size of item or a modal group presents itself. This modal size is the mode in a discrete series, while in a continuous series the mode will be located by *interpolation* in the modal group on the assumption that the frequencies of the groups on either side of the mode influence the mode in proportion to their respective numbers. Mode is usually obtained from a frequency table, and, as a working rule, by the process of regrouping. If the given series contains individual measurements, it would be better to convert it into a frequency table first.

Location of Mode: Discrete Series.

Example 1. Required to find the mode in a discrete series.

Table 5. *Location of Mode by Grouping.*

Size of item m	Frequency f					
	(1)	(2)	(3)	(4)	(5)	(6)
4	2	7		15		
5	5		13		22	
6	8	17				29
7	9		21	35		
8	12	26			**40**	
9	14		**28**			**43**
10	14	**29**		**40**		
11	**15**		26		39	
12	11	24				33
13	13		22	29		
14	9	16			20	
15	7		11			14
16	4	7				
17	3					

The frequencies given in column (1) are first grouped by two's in columns (2) and (3), and then by three's in columns (4), (5) and (6). The maximum frequency in each column is indicated in heavy type. But nno fixed point where frequency may be the largest is obtained; the mode seems to change with change in the grouping. According to column (2) it may be 10 or 11, while according to column (5) it may be 8, 9 or 10. The sizes of maximum frequency in respect of the different columns are therefore arranged, as in the following table:

Table 6. *Analysis Table.*

Column	Size of item containing maximum frequency				
(1)				11	
(2)			10	11	
(3)		9	10		
(4)			10	11	12
(5)	8	9	10		
(6)		9	10	11	
No. of times	1	3	5	4	1

From the above table we find that 10 is the size of item which is most frequented. It is not true of any other size. The mode is, therefore, located at 10.

A glance at the frequencies in column (1), table 5, might lead one to think that size 11 is the mode, since it contains the largest frequency in that column. But this impression is corrected by the process of grouping which clearly shows that mode is influenced by the frequencies of the neighbouring sizes. It is, therefore, evident that it is not always sure or correct to locate the mode by mere inspection. *It is always safer to regroup the given frequencies than to rely on inspection alone.* Inspection can give reliable results only where the frequencies run fairly regularly and mode is unquestionably clear and well-defined. We shall see it in the following example.

Location of mode: Continuous Series.

Example 2. Required to locate the mode in a continuous series whose frequency distribution is given in table 1, page 71.

We find that the frequency distribution in the said table is fairly regular and the 15-20 marks group indisputably contains the maximum frequency. Without any preliminaries of grouping we can say that the mode lies in 15-20 marks group. If we group the data, first in 10 marks group and next in 15 marks group, and then prepare from the frequency columns an Analysis Table as we did in example 1 above, we shall arrive at the same conclusion. Having known the class containing the maximum frequency we shall locate the mode in that class on the assumption already noted, *viz.*, according to the weights or influence of the neighbouring groups.

The formula for it is as follows[3]:

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}(l_2 - l_1)$$

where, Z stands for the mode[4],

l_1 and l_2 stand for the lower and upper limits of the modal group,

f_1 stands for frequencies in the modal group,

f_0 ,, ,, ,, next lower group,

f_2 ,, ,, ,, next higher group.

Applying the above formula we have,

$$Z = 15 + \frac{16 - 10}{32 - 10 - 12}(20 - 15)$$

$$= 18 \text{ marks.}$$

The process of grouping is, in fact, a method of smoothing out the irregularities of the series and can be profitably employed even in a series which does not appear to be irregular. Rather, it should be resorted to in all elementary work.

In exceptional cases, where the distribution of frequencies is very irregular, two or more groups on either side of the

---

[3] This formula, it is held, is more accurate than the customary formula, *viz.*,

$$Z = l + \frac{f_2}{f_2 + f_0}(l_2 - l_1)$$

[4] In all our further calculations mode will be represented by Z.

modal class-interval may be used as weights. But recourse should not be had to it if the series is multi-modal. It should be noted that all series do not possess a single or even a well-defined mode. Some are bi-modal, some tri-modal, *i.e.*, they have more than one mode, while some may hardly possess a mode at all. Therefore, efforts should not be wasted over forcing the appearance of an exact mode when, in fact, one does not exist in the given series. True mode should not be expected in a series which is *markedly* asymmetrical.[5] *Mode is, in fact, limited to distributions that are somewhat regular.*

Advantages of the Mode.

1. It is easily understood and has a general and precise usage.
2. It eliminates extreme (and therefore abnormal) variations. That is, its value is not affected by stray items differing much from it in their values.
3. For its determination it is not necessary to know the extreme items, except that they are few. Only the size of the middle items need be known.
4. It refers to a measurement whose expectation in the series is the greatest. That is, it is the most likely and not isolated example.
5. It can be easily located, by mere inspection in certain cases.

Disadvantages of the Mode.

1. It is frequently ill-defined and indefinite: A modal size of city may convey any meaning.
2. It is often indeterminate and, therefore, difficult to locate; this is particularly the case with irregular series.
3. It is incapable of being located by any simple arithmetical process.
4. It rejects all exceptional instances and is, therefore, not useful in those cases where weights are to be given to extreme variations.
5. It may not be fully representative of a group in which items of uniform size are comparatively small. For instance,

[5] For precise meaning of asymmetrical series see Chapter X.

if in a community of 200 people, only 4 people earned Rs. 30 each while the earnings of the rest were at any figure other than Rs. 30—and let us suppose, all the 196 people got higher than Rs. 30—and no other four people received an equal amount, then Rs. 30 would be the modal earnings although the number of people who earned more than Rs. 30 was 49 times greater than the number of those who earned Rs. 30. This appears illusory. This difficulty can, however, be got over by using class-intervals of considerable magnitude, and resorting to the process of grouping.

6. It is unsuitable for further algebraical treatment.

7. Mode multiplied by the number of items does not yield the total value of the item.

Uses of the Mode.

The concept of mode is readily intelligible, and is applied in many cases in daily routine, though involuntarily. We often hear people say 'Average calls on my telephone are 15 a day', 'Average size of shoe sold at my shop is such and such', 'An average page contains 300 words', 'An average student spends Rs. 50 a month'. In all such cases what they mean by 'average' is really mode, that is, the likeliest figure. If we are required to guess the average of a certain phenomenon, we shall generally, and rightly too, guess the mode, the dominant or prevailing size.

The use of mode is now increasing in business. It serves as a reliable guide in business forecasting. It is being realized that it is of great value in studying output. Modal output per machine can be ascertained by recording the output of similar machines and finding the output which is more or less the same over a period of time. If this modal output is left far behind in subsequent years, reasons for it may be traced back to defects in machines or their handling, lack of skill on the part of operatives, or any other inefficiency. This useful work might remain undone, or its urgency may not be felt, in the absence of a knowledge of the modal output which acts as the standard for comparison. Similarly, modal time for producing a commodity may be ascertained which would stand for the most likely time that would be required to turn out similar

goods under similar conditions. On the basis of this modal time, cost of producing a certain number of commodities may be calculated. Mode, thus, has great potentialities for being employed in business and commerce profitably.

Meteorological forecasts, which are proving very important to mercantile and other interests, are really based on the mode.

Graphical Measurement of Mode.

Mode can also be located graphically by drawing a Frequency Curve or Frequency Polygon from the given frequency distribution and reading from it the position of maximum density, which will naturally be at the apex of the curve or polygon. The method will be described in the Chapter on Graphical Presentation.

MEDIAN

Median is the value of that item in a series which divides the series into two equal parts, one part consisting of all values less, and the other all values greater than it. That is, if a series is *arrayed*, or, which comes to the same thing, the values of its items are placed side by side in ascending or descending order of their magnitude, the value of the middle item of the array is the median. If the students of a class, 43 in number, be asked to stand in order of their height, the 22nd student from either side shall be the one whose height will be called the median height of the class. This method of picking up the median item can be symbolically expressed as follows:

$$M = \text{Size of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item,}$$

where, M represents the median[6], and n the number of items.

Determination of Median: Individual Measurements.

Example 3. Required to find the median in a series of quantitative individual observations, relating to the monthly expenditure incurred by 35 students in a boarding house.

[6] Median is the *size* or *value* of the middle item, and *not the rank* or *the number* of such item.

In our further calculations M will represent the median.

The given figures are first arrayed as follows:

Table 7. *Monthly Expenditure of 35 Students arranged in Ascending order of Magnitude.*

Serial No.	Expenditure. Rs.	Serial No.	Expenditure. Rs.	Serial No.	Expenditure. Rs.	Serial No.	Expenditure. Rs.	Serial No.	Expenditure. Rs.
1	35	8	40	15	45	22	46	29	50
2	35	9	41	16	45	23	47	30	50
3	36	10	42	17	45	24	47	31	52
4	38	11	42	18	45	25	47	32	52
5	38	12	44	19	46	26	48	33	54
6	40	13	44	20	46	27	48	34	55
7	40	14	44	21	46	28	48	35	60

Applying the above formula we have,

$$M = \text{Size of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item ; } n \text{, in this case, equals 35,}$$

$$= \text{Size of } \left(\frac{35+1}{2}\right)^{\text{th}} \text{ item,}$$

$$= \text{Size of 18th item,}$$

$$= \text{Rs. } 45.$$

The number of items in the above example was odd and, therefore, there was no difficulty in locating the middle item (18th in this case). But, the number may be even. In such a case, the median is intermediate between the values of the

two middle items. Suppose, in the above example, an additional, 36th, student's expenditure was Rs. 61. Then,

$$M = \text{Size of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item ; } n \text{, in this case equals 36,}$$

$$= \text{Size of } \left(\frac{36+1}{2}\right)^{\text{th}} \text{ item,}$$

$$= \text{Size of } 18{\cdot}5^{\text{th}} \text{ item,}$$

$$= \frac{\text{Size of } 18^{\text{th}} \text{ item} + \text{Size of } 19^{\text{th}} \text{ item}}{2}$$

$$= \frac{\text{Rs. } 45 + \text{Rs. } 46}{2}$$

$$= \text{Rs. 45-8-0}$$

Determination of Median: Discrete Series.

In a discrete series also the size of $\left(\frac{n+1}{2}\right)^{\text{th}}$ item shall be the median.

Example 4. Required to locate the median of the data given in table 5, page 94.

Table 8. *Cumulative Frequency Table.*

Size of item m	Frequency f	Cumulative Frequency cf
4	2	2
5	5	7
6	8	15
7	9	24
8	12	36
9	14	50
10	14	64
11	15	79
12	11	90
13	13	103
14	9	112
15	7	119
16	4	123
17	3	126

$$M = \text{Size of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item; } n \text{ equals } 126,$$

$$= \text{Size of } \left(\frac{126+1}{2}\right)^{\text{th}} \text{ item,}$$

$$= \text{Size of } 63{\cdot}5^{\text{th}} \text{ item,}$$
$$= 10.$$

[It should be noted that in this series the size of all items beyond the 50th and up to the 64th is 10.]

Determination of Median: Continuous Series.

In the case of continuous frequency distribution, however, the median will have to be interpolated in the class-interval containing the median, if the original data from which the given frequency distribution is derived are not available. *Interpolation* gives only an approximate value. It is done on the assumption that the magnitude of the median class-interval is uniformly spread over its frequency.

Example 5. Required to locate the median in the continuous frequency distribution given in table 1, page 71.

Table 9. *Cumulative Frequency of Marks of 60 Students in Economics.*

Marks-Group m	Frequency f	Cumulative Frequency cf
0—5	4	4
5—10	6	10
10—15	10	20
15—20	16	36
20—25	12	48
25—30	8	56
30—35	4	60

$$M = \text{Size of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item ; } n \text{ equals } 60,$$

$$= \text{Size of } \left(\frac{60+1}{2}\right)^{\text{th}} \text{ item,}$$

$$= \text{Size of } 30{\cdot}5^{\text{th}} \text{ item.}$$

Item 30·5 is located in the class-interval 15—20 marks. This class is, therefore, the median-class. If we knew the value of each item in this class, the value of $30{\cdot}5^{\text{th}}$ item could have been *directly* determined. But since we don't have that information, we have to interpolate the median in the 15—20 marks-group. This group's frequency is 16 and magnitude 5. It is assumed that these 5 marks are evenly distributed over the 16 students. The 20th student gets approximately 15 marks. Therefore, the 30·5th student shall get $\left\{\frac{5}{16}(30{\cdot}5-20)\right\}$ or about 3·28 marks more than the 20th student. Thus, the size of $30{\cdot}5^{\text{th}}$ item in our series, that is the median, is 18·28 marks.

The above calculation can also be symbolically expressed as:

$$M = l_1 + \left\{\frac{l_2 - l_1}{f_1}(m-c)\right\}$$ [7]

where, M represents the median ; l_1, the lower limit of the group in which median is situated; l_2, the upper limit of median class; f_1, the frequency of median class; m, the number of middle item or $\left(\frac{n+1}{2}\right)^{\text{th}}$ item; and c, the cumulative frequency of the group lower than the one in which median is situated.

Applying the above formula we have,

$$M = 15 + \left\{\frac{20-15}{16} \times (30{\cdot}5-20)\right\}$$

$$= 15 + \left(\frac{5}{16} \times 10{\cdot}5\right)$$

$$= 18{\cdot}28 \text{ marks, approximately.}$$

[7] Formulæ slightly different from this are also given by certain authors. We, however, feel that this formula is satisfactory, as it is in keeping with the assumption we have made for interpolation.

Advantages of the Median.

1. It is easily understood.

2. It eliminates the effect of extreme (and therefore abnormal) variations.

3. It can be determined without a knowledge of the magnitude of extreme items, provided the number of items is known.

4. It is usually, *e.g.*, when found exactly, an actual example from the data.

5. It can be located by inspection in certain cases.

6. It can be exactly located (*cf* Mode, which cannot be exactly located in a multi-modal or non-modal series.)

7. It is specially useful for considering data the items of which are incapable of being quantitatively measured. A group of students may be made to stand in order of their intelligence. The middle student shall represent the median intelligence. Median can, therefore, be employed to serve as an average, yielding a sufficiently reliable representative, in an estimate of qualities like honesty, health, virtue, which cannot possibly be expressed in specific units.

Disadvantages of the Median.

1. The fact whether the median is representative of the variable depends upon the nature of the distribution of the values. It may not be representative when the distribution is irregular, *i.e.*, when the items vary greatly in magnitude. Let the runs made by the players of a cricket team be 0, 1, 3, 6, 9, 12, 48, 60, 60, 60, 98. The median runs, 12, are made only by one player, while 60 are scored by three players. Median is not a typical representative in this case.

2. It cannot be *precisely* determined when it falls between two values. Then, it can only be *estimated*. When estimated, it may be a value not found in the series. If in a class of 40 students, 20 secure marks varying between 10 and 20, and the other 20 secure marks varying from 25 to 35, the median marks would be indeterminate. They would be assumed to fall between 20 and 25, which are not obtained by any of the 40 students. The median marks would give a fictitious number.

3. It is not capable of being located by any simple mathematical process.

4. It is not useful in those cases where large weight is to be given to extreme items, for it treats all frequencies alike.

5. It is unsuitable for arithmetic or algebraic manipulation.

6. The aggregate value of items cannot be obtained when the median and the number of items are known.

7. It requires the data to be arrayed before it can be determined—an operation which involves considerable work.

Uses of the Median.

Median is easy to understand and is, therefore, useful for practical purposes. It is not only useful for the study of problems whose objects are not quantitatively measurable, but is also valuable in comparing such data as are difficult to measure individually and have to be grouped within certain limits. It is, therefore, of immense use in considering social phenomena like wages, distribution of wealth, skill, etc. The median, however, is not very suitable for being used in commerce, because commercial data are very widely dispersed, *i.e.*, they are not highly regular in distribution. Where such is the case, we have seen, median is not a good representative. What the businessman has in mind is usually the mode.

QUARTILES, DECILES & PERCENTILES

The principle according to which median is determined can be extended to divide a series into any number of parts. The values of the items dividing a series into four equal parts are called **Quartiles.** When a series is arrayed and the median divides it into two halves, each of the lower and the upper halves can also be divided into two equal parts. The value of the item dividing the lower half is called the **First Quartile** or the **Lower Quartile** represented by Q_1, and the value of the item dividing the upper half is called the **Third Quartile** or the **Upper Quartile** represented by Q_3, median being the Second Quartile. A series is, thus, divided into four equal parts at the first and the third quartiles, and the median.

Similarly, a series may be divided into ten equal parts. In doing so, we shall get nine dividing positions, the values of which are called **Deciles.** We have, thus, nine deciles in a series, the fifth decile being the median.

Again, a series may be divided into 100 equal parts, giving ninety-nine dividing positions, the values of which are called **Percentiles.** There are, thus, 99 percentiles in a series, the fiftieth percentile is the fifth decile, or the median.

In similar manner we can have Quintiles and Octiles according as our series is divided into five equal parts or eight equal parts.

Location of Quartiles, Deciles and Percentiles.

The principle of locating the median is the principle followed here also. The given series is first arrayed. Then, if the series is composed of quantitative individual observations or is a discrete one, the following formulæ shall apply:—

$$Q_1 = \text{Size of} \left(\frac{n+1}{4}\right)^{\text{th}} \text{item ;}$$

$$Q_3 = \text{Size of} \left(\frac{3(n+1)}{4}\right)^{\text{th}} \text{item ;}$$

$$D_1 = \text{Size of} \left(\frac{n+1}{10}\right)^{\text{th}} \text{item ;}$$

$$D_2 = \text{Size of} \left(\frac{2(n+1)}{10}\right)^{\text{th}} \text{item ; similarly, for the rest of the deciles ;}$$

$$P_1 = \text{Size of} \left(\frac{n+1}{100}\right)^{\text{th}} \text{item ;}$$

$$P_2 = \text{Size of} \left(\frac{2(n+1)}{100}\right)^{\text{th}} \text{item ; similarly for other percentiles ;}$$

where, Q_1 stands for first quartile[8], Q_3 for third quartile[9], D_1 for first decile, D_2 for second decile, P_1 for first percentile, P_2 for second percentile, and n for number of observations.

[8, 9] In our further calculations Q_1 and Q_3 shall stand for the first and the third quartiles respectively.

Quartiles, as also deciles and percentiles, refer to the *size* of item and not to the rank.

Individual Observations.

Example 6. In the series given in table 7, page 100.

Q_1=Size of $\left(\frac{n+1}{4}\right)^{th}$ item ; n equals 35,

=Size of $\left(\frac{35+1}{4}\right)^{th}$ item ;

=Size of 9th item,
=Rs. 41.

Q_3=Size of $\left(\frac{3(n+1)}{4}\right)^{th}$ item ; n equals 35,

=Size of 27th item,
=Rs. 48.

D_4=Size of $\left(\frac{4(n+1)}{10}\right)^{th}$ item ; n equals 35,

=Size of $14{\cdot}4^{th}$ item.

Size of 14th item+ $\frac{4}{10}$ (Size of 5th item—Size of 14th item)

=Rs. $\left[44+\frac{4}{10}(45-44)\right]$

=Rs. 44·4.

P_{90}=Size of $\left(\frac{90(n+1)}{100}\right)^{th}$ item ; n equals 35,

=Size of $32{\cdot}4^{th}$ item,

=Size of 32nd item+ $\frac{4}{10}$ (Size of 33rd item—Size of 32nd item)

=Rs. $\left[52+\frac{4}{10}(54-52)\right]$

=Rs. 52·8.

Discrete Series.

Example 7. In the data given in table 8, page 101.

$Q_1 = \text{Size of } \left(\frac{126+1}{4}\right)^{\text{th}}$ item,

$= \text{Size of } 31{\cdot}75^{\text{th}}$ item,

$=8.$

$Q_3 = \text{Size of } \left(\frac{3(126+1)}{4}\right)^{\text{th}}$ item,

$= \text{Size of } 95{\cdot}25^{\text{th}}$ item,

$=13.$

The various deciles and percentiles can also be determined in like manner.

Frequency Distributions.

If the data are grouped into certain defined limits, quartiles, deciles and percentiles shall be located by *interpolation*, which, however, yields only approximate values. The formulæ to be used are:—

$$Q_1 = l_1 + \left\{\frac{l_2 - l_1}{f_1}(q_1 - c)\right\}$$

$$Q_3 = l_1 + \left\{\frac{l_2 - l_1}{f_1}(q_3 - c)\right\}$$

where, q_1 and q_3 stand for first and third quartile numbers respectively, and other symbols for what they did in interpolating the median except that the class-intervals referred to shall be those relating to Q_1 and Q_3 and not to median. Similarly, formulæ for interpolating deciles and percentiles can be framed.

Example 8. In the continuous frequency distribution given in table 9, page 102.

$$q_1 = \left(\frac{n+1}{4}\right)^{\text{th}} \text{item} = 15{\cdot}25^{\text{th}} \text{ item.}$$

$$q_3 = \left(\frac{3(n+1)}{4}\right)^{\text{th}} \text{item} = 45{\cdot}75^{\text{th}} \text{ item.}$$

Therefore, $Q_1 = 10 + \left\{\frac{5}{10} \times (15{\cdot}25 - 10)\right\} = 12{\cdot}625$ marks,

and $Q_3 = 20 + \left\{\frac{5}{12} \times (45{\cdot}75 - 36)\right\} = 24{\cdot}0625$ marks.

Characteristics and Uses of Quartiles, etc.

Quartiles, deciles and percentiles are not averages in the sense median is, since they refer not to the whole variable but only to parts of it. Of course, for determining them the part to which they relate is treated as the whole series. Thus, quartiles are, in a sense, equivalent to the medians of the lower and the upper halves of a series, but they cannot be considered as averages of the first order, *i.e.*, as sizes which can be taken as types or substitutes of the whole series.

Yet, quartiles, etc. give valuable information regarding the series. They indicate the distance within which certain parts of the series lie. Thus, knowing the quartiles of the data given in table 7, page 100, we may say that the middle half of the series lies between Rs. 41 and Rs. 48. Deciles and percentiles can also similarly yield the information characteristic of them. We shall refer to the importance of quartiles again, while considering the manner in which items in a series are distributed.[10]

Graphical Measurement of Median, Quartiles, etc.

Median, quartiles, etc. can also be located by means of a cumulative frequency graph, generally known as the Ogive Curve, or by using what is called the Galton's method. These graphs will be taken up in the Chapter on Graphical Presentation.

ARITHMETIC AVERAGE

The arithmetic average, also called the arithmetic mean, or simply mean, is the quantity obtained by dividing the sum of the values of the items in a variable by their number. Thus, it is the average of common speech, an average quite familiar to the layman.

The term 'arithmetic average' as defined above is applied to Simple Arithmetic Average, in which all items are treated alike, each item being considered only once, as distinguished from Weighted Arithmetic Average, in which all items are *not* treated alike, each item being assigned a weight in proportion to its importance in the series.

[10] See Chapter X.

There are two methods of computing the simple arithmetic average: the direct and the short-cut.

Computing Simple Average: Individual Observations.

In following the **direct method** the values of all the items in a given series are totalled up, the total being called the Aggregate or Summation, represented by the Greek letter Σ (Capital Sigma). The quotient obtained by dividing the summation of measurements by the number of items is the required simple arithmetic average.

Symbolically.

$$a = \frac{\Sigma m}{n}$$

where, a represents arithmetic average[11], Σm represents summation of measurements, and n represents the number of items.

Example 9. Required to find the simple arithmetic average of the data given in table 7, page 100, by the direct method.

$$a = \text{Rs. } \frac{1580}{35}$$

$$= \text{Rs. } 45{\cdot}14.$$

It is not necessary that the values of the items should first be arrayed as they are done in table 7. The original data, recorded as they occurred, could have been equally well utilized to find the aggregate.

The above method of computing the simple arithmetic mean involves considerable work of addition when the series is large and digits in each number are several. To save time and labour, the **short-cut method** can be utilized if the values of the different items happen to be nearly the same. To use this method, any size of item may be assumed as the average. Deviations of the value of each item from this assumed average should then be found and put down with proper algebraic sign. The algebraic sum of these deviations divided by the number of items when added to the assumed arithmetic average gives the required simple arithmetic average.

Symbolically,

[11] In all our further calculations a will stand for the arithmetic average.

$$a-x=\frac{\Sigma d_x}{n}$$

$$\text{or, } a=x+\frac{\Sigma d_x}{n}$$

where, a stands for the arithmetic average ; x, for the assumed average ; Σd_x, for the summation of deviations from the assumed average ; and n, for the number of items.

Deviation from the assumed average is calculated by subtracting the assumed average from the size of item, that is by the formula,

$$d_x = (m-x)$$

where, d_x is the deviation from assumed average ; m, the size of item ; and x, the assumed average.

The deviation, d_x, would be positive or negative according as the size of item, m, is greater or less than the assumed average, x.

Example 10. Required to find the simple arithmetic mean of the data given below by the short-cut method.

Table 10. *Computing the Arithmetic Average by Short-cut Method.*

Size of items m	Deviations from assumed average (1362) d_x
1365	+3
1360	−2
1358	−4
1362	0
1370	+8
1363	+1
1368	+6
1364	+2
1371	+9
1362	0
	$\Sigma d_x = +23$

The algebraic sum of the deviations in the above table is + 23. Then, according to the formula we have,

$$a - 1362 = \frac{23}{10}$$

$$\therefore \quad a = 1364{\cdot}3.$$

The above short-cut method is based on the simple fact that the **algebraic sum of the deviations of individual values from the arithmetic average is zero.** Thus, deviations of the size of items from 1364·3 in the above example are respectively,

$$+{\cdot}7, -4{\cdot}3, -6{\cdot}3, -2{\cdot}3, +5{\cdot}7, -1{\cdot}3, +3{\cdot}7, -{\cdot}3, +6{\cdot}7, -2{\cdot}3.$$

Their summation is zero. This also illustrates the Law of Averages, mentioned earlier.

Computing Simple Average: Discrete Series.

In a series of discrete type each size of item should first be multiplied by its frequency, and the product summated and divided by the total frequencies. The quotient would give the simple arithmetic average according to **direct method.**

For the reason for which we used it in example 10 above, the **short-cut method** can also be employed in a discrete series. First, deviations of each size of item from the assumed average should be found out as in the above example. Each deviation should be multiplied by its frequency and algebraic sum of the products obtained. Then, the same short-cut formula as used in the above example shall yield the required simple arithmetic average. Example 11 shall demonstrate the working of these two methods in a discrete series.

Example 11. Required to find the simple arithmetic average of the data given in table 8, page 101, by the direct and the short-cut methods.

Let x, assumed average, for the latter method be 10.

Table 11. *Calculation of Simple Arithmetic Average by the Direct and the Short-cut Methods.*

1	2	3	4	5
Size of items	Frequency	Total size of items (col. 1× col. 2)	Deviations from assumed mean (10)	Total Deviations (col. 2× col. 4)
m	f	mf	d_x	fd_x
4	2	8	−6	−12
5	5	25	−5	−25
6	8	48	−4	−32
7	9	63	−3	−27
8	12	96	−2	−24
9	14	126	−1	−14
10	14	140	0	0
11	15	165	+1	+15
12	11	132	+2	+22
13	13	169	+3	+39
14	9	126	+4	+36
15	7	105	+5	+35
16	4	64	+6	+24
17	3	51	+7	+21
	$n=126$	$\Sigma m=1318$		$\Sigma d_x=+58$

In computing a by the direct method we shall be concerned with columns (1), (2) and (3) ; while in calculating it by the short-cut method we shall be concerned with columns (1), (2), (4) and (5).

Direct method : $a=\frac{\Sigma m}{n}=\frac{1318}{126}=10{\cdot}46.$

Short-cut method: $a-x=\frac{\Sigma d_x}{n}$

$\therefore\ a=10+\frac{58}{126}=10{\cdot}46.$

Computing Simple Average: Continuous Series.

When frequency distribution of a continuous type is given arithmetic average can only be calculated on the assumption that the values of all the items in each class are identical with the mid-value of the class-interval. Both the direct and the short-cut methods can be followed. After finding out the mid-values of the class-intervals, the procedure of calculating the arithmetic average would be the same as that in the case of discrete series.

Example 12. Required to find the simple arithmetic average of the data given in table 1, page 71, by the direct and the short-cut methods.

In table 12, we shall not be concerned with columns (5) and (6) for the direct method, and with column (4) for the short-cut method.

Table 12. *Calculation of Simple Arithmetic Average of Marks of 60 Students by the Direct and the Short-cut Methods.*

1	2	3	4	5	6
Marks group.	Mid-value m	Frequency f	Total value of items (col. 2 × col. 3) mf	Deviations from assumed mean (17·5) d_x	Total deviations (col. 3 × col. 5) fd_x
0—5	2·5	4	10	−15	−60
5—10	7·5	6	45	−10	−60
10—15	12·5	10	125	− 5	−50
15—20	17·5	16	280	0	0
20—25	22·5	12	270	+ 5	+60
25—30	27·5	8	220	+10	+80
30—35	32·5	4	130	+15	+60
		$n=60$	$\Sigma m=1080$		$\Sigma d_x=+30$

Direct method : $a=\frac{\Sigma m}{n}=\frac{1080}{60}=18$ marks.

Short-cut method : $a-x=\frac{\Sigma d_x}{n}$

$\therefore \; a=17{\cdot}5+\frac{30}{60}=18$ marks.

In the foregoing examples on simple arithmetic average both the direct and the short-cut methods have been demonstrated. It will be seen that the answers in each case by both the methods are exactly the same. The saving in labour resulting from the use of the short-cut method is quite obvious. In the above examples, it is not necessary to use as assumed averages only those values which we have chosen for the purpose. Other values may also be so used. The answers will not be different.

Advantages of Simple Arithmetic Average.

1. It is easily understood and has a general usage.
2. It is easy to calculate. Its calculation is a common knowledge.
3. It utilizes all the data in the group.
4. It does not necessitate the arraying of data as the median does, nor the grouping of data as the mode does.
5. It can be known even when number of items and their aggregate values are known, and details of the different items are not available.
6. It is determinate. It is not indefinite.
7. The aggregate can be calculated if the number of items and the average are known.
8. It affords a good standard of comparison, since the abnormalities in opposite directions tend to cancel each other if the number of items is sufficiently large.
9. It is amendable to algebraic and arithmetic manipulation.

For all these qualities it is the most widely used average.

Disadvantages of Simple Arithmetic Average.

1. It may give considerable weight to extreme (and therefore abnormal) items. A millionaire would greatly affect the

average income of a town where the majority consists of ill-paid artisans.

2. It can hardly be located by inspection ; mode and median can be.

3. It can ignore any single item only at the risk of losing its accuracy. Mode and median can be computed even when the values of extremes are not known.

4. The average that results may not occur in the data at all, and may not, therefore, be representative to the fullest degree. The average of 2, 4 and 9 is 5, which does not occur in the series.

5. It cannot be used when the data are incommensurable: median can be used in qualitative studies.

6. This average might lead to fallacious conclusions when the actual figures from which it is obtained are not given. For instance, two students, A and B, get the following marks:—

		A	B
First Terminal Exam.	..	40%	60%
Second Terminal Exam.	..	50%	50%
Annual Examination	..	60%	40%

The average percentages of both of them are identical, 50. But A's progress is positive while B's is negative. If the average, 50, is not supported by the percentage marks in the three examinations, the fact that A is progressing while B is deteriorating would be concealed and a fallacious conclusion that the standard of both of them is the *same would* be drawn.

Uses of Simple Arithmetic Average.

It is used in many social and economic studies. Its use is daily routine in business and commerce. It is an average which even a 'man in the street' understands. It is the common average. Statistics uses it not only as a type for comparison, but for several other statistical calculations as well. "Average output of a commodity," "Average imports or exports over a period," "Average cost of production" "Average price"—in all such expressions the average used is generally the arithmetic average.

WEIGHTED ARITHMETIC AVERAGE

In computing simple arithmetic average it was assumed that all items were of equal importance. This may not always be the case. Where items vary in importance they must be assigned weights in proportion to their relative importance. The value of each item is then multiplied by its weight, products summated and divided by the number of weights and *not* by the number of items. The quotient is the weighted arithmetic average.

Symbolically,

$$a' = \frac{\Sigma\ (wm)}{\Sigma\ (w)}$$

where, a' stands for weighted arithmetic average ; w, for weight ; and, m, for measurement.

Weights may be actual or approximate. Actual weights should be used where they are available ; otherwise, they may be estimated on the strength of the best possible data available. For instance, if we know the *actual number of people engaged* on the teaching, clerical and menial staff of an institution and the *average earnings* of each class of employees, we should multiply the average earnings of each class by the actual employees in the corresponding class, summate the products and divide the sum by the number of employees to secure the weighted arithmetic average of earnings. The actual number of employees in each class shall constitute the actual weight of the class. If, however, information about the actual number of employees in each class is not available, the best possible estimate of the proportionate strength of the different classes may be made and used as weights. The following example illustrates the procedure of working out the weighted mean.

Computation of Weighted Average: Direct Method.

Example 13. Required to compute weighted arithmetic average from the data given in table 13, on next page, by the direct method.

Table 13. *Calculation of Weighted Average Earnings of the Employees of X College.*

1	2	3	4	5	6
Description of the employees	Number of employees w	Monthly Average Earnings m	Product of columns (2) and (3) wm	Estimated Weights w'	Product of columns (3) and (5) $w'm$
		Rs.	Rs.		Rs.
Professors ..	2	600	1,200	1	600
Lecturers ..	16	200	3,200	8	1,600
Demonstrators ..	4	100	400	2	200
Clerks ..	2	60	120	1	60
Peons ..	7	15	105	4	60
Watchmen ..	3	14	42	1	14
Total ..	$\Sigma w=34$	$\Sigma m=989$	$\Sigma wm=5{,}067$	$\Sigma w'=17$	$\Sigma w'm=2{,}534$

Weighted arithmetic average—

(*a*) by using actual weights $= \dfrac{\Sigma\,(wm)}{\Sigma\,(w)}$

$= \text{Rs. } \dfrac{5067}{34}$

$=$ Rs. 149-0-6.

(*b*) by using estimated weights $= \dfrac{\Sigma\,(w'm)}{\Sigma\,(w')}$

$= \text{Rs. } \dfrac{2534}{17}$

$=$ Rs. 149-0-11.

What is Weighting ?

Along with demonstrating the method of computing the weighted average, the above example also shows that it is not necessary that the weights applied should be actual ones (as in column 2) ; they may be approximate also (as in column 5), the difference between the results obtained by using the actual and estimated weights being only five pies, which is not material. It should, however, be noted that when the number of weights used is small, their size may have a considerable effect upon the average, and therefore, if *estimated weights are used they should be approximately correct.* If many weights are used, the error in their estimation will be mostly unbiassed and, therefore, cancelling one another. The average would not, then, be materially affected.

When we desire to calculate the average earnings per employee in example 13, above, they might, at first sight, appear to be $\frac{600+200+100+60+15+14}{6}$ or Rs. 164·6. If it were so, the total monthly earnings would be Rs. (164·6×34) or Rs. 5596·4; but, in fact, the monthly pay roll amounts to (Rs. 600×2) +(Rs. 200×16) + (Rs. 100×4) + (Rs. 60×2) + (Rs. 15×7) + (Rs. 14×3), that is, Rs. 5067 only. Since Rs. 164·6 multiplied by the number of employees does not yield the monthly pay roll, it is not the correct average. If, however, we multiply the weighted average, Rs. 149-0-6, by the number of employees, 34, we get Rs. 5067, the monthly pay roll. The weighted average of the average earnings, in this case, is approximately the same as the simple arithmetic average of the total earnings of all the 34 employees would be. In other words, instead of totalling up the earnings of 34 employees, we have counted Rs. 600 two times, Rs. 200 sixteen times, Rs. 100 four times and so on, and summed up the resulting products. Thus, 'weight' indicates the number of times a given item is to be counted, and is therefore equivalent to frequency. The sum of the weights, like that of frequencies, is thus equal to n.

The multiplication, in the above example, of the size of items (monthly average earnings) by their corresponding frequencies (number of employees) has been called weighting by

some writers[12] on statistics, while it appears to be another device for computing the simple arithmetic average, as told above. A few writers therefore do not regard it as weighting and their view is justified. Horace Secrist, for instance, is of the opinion that weights should be 'determined by some evidence of importance other than that associated with the items themselves'.[13] Similarly, to Kelley weights are 'determined not at all, or not solely, by the population, but from other evidences of importance'.[14] Thus, suppose in a factory there are 30 employees in all, 18 of whom are male adults, 8 female adults and 4 children, and further suppose their relative productivity is 1, 0·75 and 0·5 respectively, the men equivalents of these 30 employees will be (18+6+2)=26 only. In this case, relative productivity will constitute the weights.

Similarly, a teacher may assign weights to different grades of work done by the students in proportion to the importance of the grades. He may, for example, assign 4 to seminar work, 2 to class-room work and 3 to monthly test. Marks obtained out of, say, 100 in each grade will be multiplied by the weight of the grade and the sum of the products divided by 9. This average will not correspond closely to the simple average as shown in example 14, below. Likewise, a wholesale price index number may be weighted in proportion to the marketable surpluses of different commodities, or values thereof, and a cost of living index number may be weighted by the proportions which expenses on different items of consumption bear to total expenditure. Weights are thus numbers which stand for the relative importance of items in a variable.

Computation of Weighted Average: Shortcut Method.

In the short-cut method, any size of item may be assumed as the average, deviations of each item from this average computed and multiplied by the connected weight, the products summated and divided by the sum of weights, the quotient being added to the assumed average to arrive at the required weighted average.

---

[12] Notably, King, Boddington, Connor.

[13] Horace Secrist, *An Introduction to Statistical Methods*, New York, 1933, p. 280.

[14] Kelley, T. L., *Statistical Method*, New York, 1923, p.68.

Symbolically,

$$a'-x'=\frac{\Sigma\,(wd_x')}{\Sigma\,(w)}$$

where, a' stands for the weighted average ; x' for assumed weighted average ; w for weight and, d_x' for deviation from assumed weighted average.

Example 14. Required to compute weighted arithmetic average of the data given below by the short-cut method.

Table 14. *Calculation of Weighted Average Marks of a Student.*

Examination	Marks out of 100 m	Deviation from assumed average (50 marks) d_x'	Weight w	Product of deviation and weight wd_x'
Seminar ..	45	− 5	4	−20
Class-work ..	62	+12	2	+24
Monthly Test ..	52	+ 2	3	+ 6
	$\Sigma m=159$		$\Sigma w=9$	$\Sigma wd_x'=10$

Let the assumed weighted average be 50 marks. Then,

$$a'-x'=\frac{\Sigma\,(wd_x')}{\Sigma\,(w)}$$

$$\text{or}\quad a'-50=\frac{10}{9}$$

$$\therefore\quad a'=51\%\ \text{marks, approx.}$$

The simple arithmetic average in the above example will not approximate closely to the weighted average. For,

$$a=\frac{\Sigma m}{n}=\frac{159}{3}=53\%\ \text{marks.}$$

In fact, both the systems—allotting weights according to actual number and according to estimates of relative importance—are in vogue. We shall read more of them while discussing Index Numbers.

When Should Weighted Average Be Used ?

(1) *When the items falling into different grades or classes of the same group show considerable variation, and it is desired to obtain an average which would be representative of the whole group, weighted average is the only proper average to be used.*

It is usually found in a study of wages that the number of workmen earning high wages is much less than the number of those getting low wages. If, then, a simple arithmetic average of the wages in all the occupations—treating all grades as of equal importance—were computed, the wage of the manager would be given as much weight as that of a coolie or a gang-man and the average would appear considerably large. Weights cannot be ignored in such cases.

(2) *When the averages of the different grades constituting a group of observations and the number of observations falling in each grade are known, the only proper way to compute an average of the averages is to strike their weighted average.* Consider the following:—

I series: 2, 4. (Two items only).
II series: 5, 6, 7. (Three items only).

Simple average of I series $= \dfrac{2+4}{2} = 3$

Simple average of II series $= \dfrac{5+6+7}{3} = 6$

Simple average of the two series combined

$$= \frac{2+4+5+6+7}{5} = 4{\cdot}8$$

Simple average of the two averages $= \dfrac{3+6}{2} = 4{\cdot}5$

Weighted average of averages of the two series

$$= \frac{(3\times2)+(6\times3)}{5} = 4{\cdot}8$$

The simple average of the two averages, 4·5, no doubt gives an idea of the general average of the two series combined but the weighted average of averages, 4·8, is a better representative. It also coincides with the simple average of the two series combined.

The above principle also underlies example 13, solved above, where the weighted average of average earnings of the different classes of employees was found to be true average.

(3) *Weighted average should also be used when the size of items changes and the relative proportion of the number of items also changes.* For instance, if, in example 13, the number of employees in each grade were doubled or trebled so that the relative proportion of the number of employees remained unaffected, the average already computed, *viz.*, Rs. 149-0-6 would stand as representative even after the doubling or trebling of employees. But, if their number changed such that their relative proportion were also changed, a new weighted average will have to be computed.

(4) *When percentages relating to the different grades of a group of observations are given, weighted average of the percentages alone will give the correct average for the group.* If the percentages of success at the M.A., M.Sc., B.A., B.Sc., and B.Com. examinations of a university in a particular year were 80, 70, 65, 60 and 75 respectively, the average percentage success for the five examinations would not be the simple arithmetic average of the five percentages, which is 70. These percentages will have to be weighted by the number of candidates that appeared at the different examinations if the number is known, otherwise by the best estimate available thereof. If out of a total of 250, the relative strength of the candidates were 20, 10, 100, 80 and 40 for the five examinations in question, the weighted average of the percentages would be 66·4.

This result is also corroborated by the fact that the given percentages when applied to the respective strength of the different examinations would show that 166 out of every 250 candidates came out successful in the five examinations taken together.

Caution in Weighting.

Weights should be used with caution, since proper weighting is as valuable as manipulated or erroneous weighting is dangerous. Weights should be as approximately accurate as possible and should be determined in the light of the best evidence available. We take below an example to demonstrate this principle.

Example 15. Required to compute weighted arithmetic average of the percentage success in two universities.

Table 15. *Calculation of Weighted Average of the Percentage Success in X and Y Universities.*

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
University Examination	Relative proportion of candidates	Percentage of success in X university	Product of columns 2 & 3	Percentage of success in Y university	Product of columns 2 & 5	Arbitrary weights	Product of columns 3 & 7	Product of columns 5 & 7
1. M.A.	10	80	800	75	750	15	1200	1125
2. M. Sc.	7	65	455	50	350	10	650	500
3. B.A.	40	60	2400	70	2800	10	600	700
4. B. Sc.	25	55	1375	75	1875	5	275	375
5. B. Com.	13	75	975	65	845	40	3000	2600
Total	95	335	6005	335	6620	80	5725	5300

I Simple average of percentage success in—

$$X \text{ University} = \frac{335}{5} = 67$$

$$Y \text{ University} = \frac{335}{5} = 67$$

II Weighted average, by using the weights in column 2, in—

$$X \text{ University} = \frac{6005}{95} = 63{\cdot}2$$

$$Y \text{ University} = \frac{6620}{95} = 69{\cdot}7$$

III Weighted average, by using the weights in column 7, in—

$$X \text{ University} = \frac{5725}{80} = 71{\cdot}6$$

$$Y \text{ University} = \frac{5300}{80} = 66{\cdot}3$$

The above example makes the following facts clear :—

(i) The simple averages for both the universities are identical. If they are used for comparing the percentage success, it would appear that the average percentage success in both of them is the same.

(ii) But when weights are assigned to the results in proportion to the number of candidates at different examinations, Y university appears to have much better result than X. This conclusion is corroborated by the fact that in the B.A. and B.Sc. examinations, where the number of candidates is very large, the percentage success in Y university is higher—much higher in B.Sc.—than in X. These weighted averages, therefore, yield a proper comparison.

(iii) When weights given in column 7, chosen quite arbitrarily without regard to the proportionate importance of the items, are used, the conclusion arrived at in (ii) above is reversed: X university appears to indicate better results than Y. This

inference is not corroborated by facts. Therefore, allotting of weights needs great care and caution lest fallacious conclusions result.

(iv) A comparison of the averages arrived at by using the weights given in columns (2) and (7) reveals that the averages for X university in II and III cases, 63·2 and 71·6, and also the averages for Y in the two cases, 69·7 and 66·3, are very much different between themselves. It is so, because the weights in column (7) do not have between one another the same proportion as weights in column (2) do. In our example of the earnings of the employees of X college, table 13, the weights in column (5) have the same proportion among themselves as those in column (2) do. The averages resulting from the use of the two weights are, therefore, not materially different. It is thus established that **it is not the absolute size of the weights but their relative size that is important. A weighted average is unaltered if all the weights are multiplied or divided by the same quantity,** that is, if their mutual proportions remain unchanged. Even if the multiplication or division is not accurate but only approximate, as in column (5), table 13, the resulting weighted average shall be approximately correct.

Difference between Simple and Weighted Averages.

Strictly speaking the so-called "unweighted" (simple) average is a special case of "weighted" average in which the weights are unity for each item. In the "weighted" average, weights are chosen according to some systematic plan and are not unity for each item.

However, in example 13, the simple average was Rs. 164·6 while the weighted average was Rs. 149-0-6. Similarly, in the example considered on pages 122-123 the simple average of 3 and 6 was 4·5 whereas their weighted average was 4·8. It is thus established that any weighted mean will generally differ from the unweighted average of the same quantities. This difference, in general terms, may be expressed as follows :

If the weights are so distributed that the larger items are associated with smaller weights, and the smaller items with larger weights, the smaller items will have greater influence in the formation of the weighted average and the weighted average will be less than the simple average. If, on the other hand, larger items have larger weights and the smaller ones have smaller weights, the larger items will have more influence and the weighted average will be more than the simple average. But if there is apparently no connection between weights and items, that is, if larger weights are as much associated with larger as with smaller items, and similar is the case with smaller weights, the difference between the weighted and the simple averages may be insignificant.

Thus, the simple average of five items, 1, 2, 3, 4 and 5, is 3 ; their weighted average, if their weights be, respectively, 5, 4, 3, 2 and 1, is 2·33, which is less than the simple average ; their weighted average, if their weights be, respectively, 1, 2, 3, 4 and 5, is 3·66, which is more than the simple average ; and, their weighted average, if their weights be, respectively, 3, 5, 1, 2 and 4, is 2·93, which is not materially different from the simple average.

This appreciation of the difference between the simple and weighted averages is of great value in certain cases. For instance, if we have reason to believe that weights are not connected with the size of items in a particular way, that is, as likely as not large and small weights are attached indiscriminately to large and small sizes of items, we shall not be unjustified in taking the simple average as a fairly good approximation to the unknown weighted average. We shall then avoid the strain involved in discovering exact weights.

Characteristics and Uses of Weighted Average.

The weighted average, generally speaking, possesses the same advantages and disadvantages and has almost the same uses as the simple arithmetic average. Like the latter, it gives the characteristics of the whole group to which it pertains and does not refer to the actual conditions of any one grade or class

of a group. It is useful for general comparison. It is a good average to use when a group as a whole, say the whole of an industry, is surveyed. If, however, we wish to study the actual condition of the individual grades or classes of a group, we should compute the averages of the different grades separately and compare such averages. That is, we should study each homogeneous part separately.

The weighted averages is invariably applied in the calculation of birth, marriage and death rates, and their comparison in different places or at different times. Its application to death rates will be treated towards the end of this Chapter, and to index numbers in the Chapter on Index Numbers.

It may be useful to remember that weighting is essential for attaining accuracy in the result when the series is small; it is unnecessary in a very long series since in such series the weighted and simple averages tend to be identical.

GEOMETRIC MEAN

The geometric mean, also called the geometric average, is the *n*th root of the product of the *n* quantities of a series. The geometric mean is obtained by multiplying the values of the items together and extracting the root of the product corresponding to the number of items. Thus, the square root of the product of two quantities is their geometric mean. Similarly, the cube root of the product of three quantities is the geometric mean of three quantities. Symbolically,

$$g=\sqrt[n]{a\times b\times c\times \ldots\ldots n}$$

where, g stands for the geometric mean; n, for the number of items and $a, b, c \ldots\ldots n$, for the values of n items.

The geometric mean of 4 and 9 is equal to $\sqrt{4\times9}=6$; the geometric mean of 2, 4 and 8 is equal to $\sqrt[3]{2\times4\times8}=4$. When the number of items in a series is larger than three, this process is difficult to follow. To obviate the difficulty, logarithm of each size is obtained from a Mathematical Table.[15] The

[15] Mathematical Tables are given at the end of the book.

logarithms of all the values are added up and divided by the number of items. The anti-logarithm of the quotient is the required geometric mean. The formula is :

$$g = \text{Anti-log}\left[\frac{(\log a + \log b + \log c + \ldots . \log n)}{n}\right]$$

The geometric mean is thus the anti-log of the average logs of the numbers.

Example 16. Required to calculate the geometric average of the data given below in the first column.

Table 16. *Calculation of Geometric Average.*

Size of Items	Logarithms
4·5	0·6532
250·5	2·3979
12·0	1·0792
119·5	2·0792
30·0	1·4771
42·0	1·6232
75·0	1·8751
35·4	1·5490
Rs. 568·4	12·7339

According to the formula we have,

$$g = \text{Anti-log}\ \frac{12{\cdot}7339}{8}$$

$$= \text{Anti-log } 1{\cdot}6,$$
$$= \text{Rs. } 39{\cdot}81$$

The geometric mean of a variable is always less than its simple arithmetic average, unless all the sizes of the variable

are equal in magnitude in which case the geometric and the arithmetic means will be identical. In the above example,

$$a = \text{Rs. } \frac{568{\cdot}4}{8} = \text{Rs. } 71{\cdot}05,$$

which is greater than the geometric mean.

WEIGHTED GEOMETRIC MEAN

To compute the weighted geometric mean of a series of items, each individual item should first be multiplied by itself a number of times equal to its weight; the products so obtained should then be multiplied by each other. The nth root of this final product, where n stands for the total number of weights, is the required weighted geometric average. Symbolically,

$$g' = \sqrt[n]{a^{w_1} . \ b^{w_2} . \ c^{w_3} \ldots\ldots . n^{w_n}}$$

where, g' represents weighted geometric mean, $a, b, c, \ldots\ldots n$ represent the values of n items, $w_1, w_2 \ldots\ldots w_n$ represent the weights corresponding to the size of item to which they relate, and n represents the total number of weights.

In practice, logarithms may be used. First, logarithm of each individual item should be found from a mathematical table. Each log should then be multiplied by its weight. The summation of such products divided by the total number of weights is the required weighted geometric mean. This may be expressed as follows :

$$g' = \text{Anti-log.} \left[\frac{(\log a \times w_1) + (\log b \times w_2) + \ldots\ldots (\log n \times w_n)}{w_1 + w_2 + \ldots\ldots w_n} \right]$$

Example 17. Required to compute weighted geometric mean of the data given below.

Table 17. *Calculation of "Capital" Index of Indian Industrial Activity, March, 1942, Based on Weighted Geometric Mean.*

Items	Weights	Index No. (1935=100)	Log.	Weight×Log.
Indian Cotton Consumption ..	9	149·5	2·1761	19·5849
Jute Manufacturers ..	6	134·9	2·1303	12·7818
Steel Ingots ..	5	147·1	2·1673	10·8365
Pig Iron ..	8	134·4	2·1271	17·0168
Paper ..	3	185·2	2·2672	6·8016
Coal ..	7	110·0	2·0414	14·2898
Rail and River-borne Trade ..	24	108·9	2·0374	48·8976
Cheque Clearances ..	20	88·5	1·9469	38·9380
Notes in Circulation* ..	6	132·4	2·1206	12·7236
Consumption of Electricity ..	7	152·9	2·1847	15·2929
	95			197·1635

According to the formula we have,

$$g' = \text{Anti-log.} \frac{197{\cdot}1635}{95}$$

$$= \text{Anti-log. } 2{\cdot}0754$$

$$= 118{\cdot}9$$

The "Capital" Index of Indian Industrial Activity for March 1942, therefore, is 118·9.

Advantages of the Geometric Mean.

1. It is determinate, provided the values of the variable are greater than zero.

* April, 1935 to March, 1936=100.

2. It is based on all the data in the group.

3. It gives less weight to large items and more to small ones than does the arithmetic average.

4. It is particularly useful when dealing with ratios.

5. It is amenable to arithmetic and algebraic manipulation.

Disadvantages of the Geometric Mean.

1. It cannot be used when any of the quantities is zero or negative ; for, when a quantity is zero, the product of all quantities will be zero and the g will be zero, and when a quantity is negative, the product of all quantities will be negative and the g will become unrepresentative and imaginary.

2. It may be found to lie at a point where very few (or even none) of the actual measurements lie.

3. It entails much work of calculation and is difficult of computation.

4. It is less easily understood than the arithmetic average.

Uses of the Geometric Mean.

The property of giving large weight to small items makes geometric average a very suitable type in studying various social and economic phenomena where it is desired to give large weight to small items. If some items in a series are very big and others very small it is not the arithmetic average, median or mode but the geometric mean that yields a representative type. If the annual incomes of, say, the employees of a university vary between Rs. 180 and Rs. 24,000, geometric mean of the incomes will give a good idea of their average yearly income. If arithmetic average were used, a single salary of Rs. 24,000 would pull the arithmetic mean very high, because of the comparatively very low salaries of clerks and peons. The geometric mean would nullify the effect which large values have upon the arithmetic average. It may be remembered that if in a series the arithmetic and the geometric means are found to differ considerably from each other, the geometric mean should be regarded as better representative of

the two, since it falls within a range of the majority of the given examples.

Another important use of geometric mean is in connection with index numbers. Index numbers are ratios, and the geometric mean is particularly useful in dealing with relative as against absolute differences.

Geometric mean is used in the construction of the Index Numbers of Wholesale Prices issued by the Economic Adviser to the Government of India. It is used in the *Board of Trade* Index Number of Wholesale Prices in Great Britain. It was used by Professor W. S. Jevons in his study of the changes in the general level of prices. The difficulty experienced in its calculation and the fact that it is too abstract to be readily intelligible have stood in the way of its popularity and general use. It is, however, useful in the averaging of ratios and rates.

HARMONIC MEAN

The Harmonic Mean, also called the harmonic average, is the total number of items of a variable divided by the sum of the reciprocals of the values of the variable. Symbolically,

$$h = \frac{n}{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \ldots\ldots\ldots \frac{1}{n}}$$

where, h stands for the harmonic mean ; $a, b, c \ldots . n$ represent the values of the n items of the variable, and n is the number of items. Reciprocals of numbers can be easily obtained from a Mathematical Table.[16]

The harmonic mean can also be expressed as the reciprocal of the arithmetic average of the reciprocals of the values of the items of a series. Thus,

$$h = \text{Reciprocal } \frac{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \ldots\ldots\ldots \frac{1}{n}}{n}$$

[16] The Mathematical Tables given at the end of the book give reciprocals of natural numbers.

Example 18. Required to calculate the harmonic mean of the data given in table 16, page 129.

Table 18. *Computation of Harmonic Mean.*

Sizes of Items Rs.	Reciprocals.
4·5	·2222
250·0	·0040
12·0	·0833
119·5	·0084
30·0	·0333
42·0	·0238
75·0	·0133
35·4	·0283
	·4166

According to the formula we have,

$$h=\frac{8}{\cdot 4166}=\text{Rs. } 19{\cdot}2$$

$$\text{or, } h=\text{Reciprocal }\frac{\cdot 4166}{8}$$

$$=\text{Reciprocal }\cdot 05208$$
$$=\text{Rs. } 19{\cdot}2.$$

The harmonic mean is always less than the geometric mean of the same items. In the above example :

a=Rs. 71·05
g=Rs. 39·81
h=Rs. 19·2

WEIGHTED HARMONIC MEAN

Weighted harmonic mean is calculated in the same manner as the weighted geometric mean ; that is, the reciprocals of the

given items are multiplied by the weights, and the products summated and averaged. The reciprocal of the resulting average is the weighted harmonic mean of the given series.

Characteristics and Uses of Harmonic Mean.

The harmonic mean is determinate and considers the values of all items of the data. It gives the largest weight to the smallest items, and is valuable where such weighting is desirable. It may be used in averaging of rates and time. It is used in very special cases and is not suitable for general application. The time and trouble involved in its calculation also stand in the way of its popularity. It is abstract and not easy to understand. It may not be an actual item occurring in the series.

SOME IMPORTANT CONSIDERATIONS

Averages of the First Order.

The various averages discussed so far are averages of the "first" order—that is, they deal with the actual values of a statistical variable. In contrast to them, we shall later (in Chapter X) study averages of the "second" order—that is, those which summarize not the actual values of items but the difference between such values and their average. Averages of the "first" order can be used as representatives or substitutes of the data to which they relate.

Typical and Descriptive Averages.

It is always arithmetically possible to calculate an average from a given series. But, this does not imply that every average is statistically significant. If the average of a series is found to lie near a point round which the data exhibit a tendency to cluster, the average may be presumed to be sufficiently representative of the series. It is then called a "Typical Average". If, on the other hand, the distribution of items is irregular so that the data seem to cluster round several points or do not cluster at all, the average has only arithmetical significance and should not be considered as fully representative of the series. It is then called a "Descriptive Average". Typical average can be substituted for the series for purposes of comparison or for other information relating to the series.

Relative Position of Averages.

As shown in example 18, above, the arithmetic average is greater than the geometric mean of the same quantities, and the geometric mean is greater than their harmonic mean. These three means will be identical if all the sizes of items in the given series are of equal value. Symbolically, this may be expressed by the inequalities

$$a \geqq g \geqq \text{h},$$

where the sign of equality prevails only when all the sizes are equal.

If a given group is symmetrical, the arithmetic mean, mode and median are coincident. In a moderately asymmetrical distribution the following relationship between the mean, mode and median may be found to hold good :

$$(M-\text{Z}) = \tfrac{2}{3}\ (a-\text{Z})$$

This formula may be used to find the value of one average if the values of the other two averages are given, *provided* the series to which the averages refer is moderately asymmetrical. Graphically, the above relationship may be studied from Figure 2 in Chapter X. It may be remembered that if there is a small group at a high degree, the arithmetic mean would probably be above the median and the mode below it. The arithmetic average would probably be below the median and the mode above it, if there is an absence of high numbers. It should be clearly noted that these rules are only tentative.

Choice and Properties of Average.

An average is a summarized expression of a series of divergent individual values. All averages do not characterize the series in the same way. They yield only that information which, by their nature, they are able to transmit. This information differs according to the kind of average used. Therefore, *it is the purpose for which an average is to be employed that will largely determine its choice.* No one average is good for all purposes. The mean chosen to study a given phenomenon should be such as will throw light on the striking features and the essential characteristics of the phenomenon. Different averages will apply to different groups of data ; each must be selected on its own merits. Each average is affected

differently by the distribution, frequencies and the character of the details. A knowledge of its preculiarities or characteristics is, therefore, a pre-requisite for scientific use of an average. It is evident that an average simplifies complexity, but if the particular merits, demerits, scope and characteristics of the average are neglected, the simplicity arrived at shall not be worth having. Caution, foresight and analysis are, therefore, necessary in the use of averages. If they are ignored, the very principles on which scientific method rests shall be violated. This is not desirable.

What are the desirable properties for an average to possess ? First, it should be rigidly defined ; second, it should be based on all the observations of the data ; third, it should be readily comprehensible ; fourth, it should be capable of being computed with reasonable ease and rapidity ; fifth, it should be as little affected by fluctuations of sampling as possible ; and last, it should be readily amenable to arithmetic or algebraic treatment.

From a perusal of the advantages and disadvantages of the various averages outlined in the foregoing pages, it will be evident that the arithmetic average—the common mean—possesses the above properties more than any other single average does. It is rigidly defined, is based on all observations, is readily comprehensible, is less affected by fluctuations of sampling than, say, the median, and above all, is suitable for algebraic treatment. Of course, median is somewhat more easily computed than the arithmetic average, but median is often indeterminate and its algebraic treatment is difficult, if not impossible, in many cases. Mode is hardly useful in elementary work owing to the difficulty of locating it with precision. Since the arithmetic average uses all the items of a given series, it is likely to be less erratic, *i.e.*, less sensitive to small change in values of individual items. *The arithmetic average is, therefore, quite suitable for all general purposes, unless there is special reason to select any other average.* For instance, if items of small values are far larger in a series than items of large values, arithmetic average will not be a good average to use ; geometric mean will be better. For measuring periodic increase in, say, population it is not the

arithmetic, but the geometric mean that will be more suitable. If it is a question of measuring the average ratio of change in, say, prices over a period of time, geometric mean will be the appropriate average to use, since it gives equal importance to equal ratios of change (see Chapter on Index Numbers). If in the given series it is necessary to give more weight to the smallest items than to others, harmonic mean will be the proper average to compute. If a train runs 20 miles per hour for the first mile and 30 miles per hour for the second mile, the average speed of the train would not be the arithmetic average of the two speeds (*viz.*, 25 miles per hour) but the harmonic mean, that is 24 miles per hour. Similarly, if an enquiry is made into the 'average' size of shoe sold at a shop or an 'average' coat tailored at a tailor's, it is not the arithmetic average but mode that will serve the purpose. Again, if an idea of average intelligence of a class is to be had in the absence of any quantitative measure of intelligence, median shall be the only average to use. Where the extreme values in a series are to be ignored, median and not the arithmetic average will constitute the typical average ; as, for example, in computing the average price of a commodity for a particular month for a State when districtwise quotations are available. These are typical cases where arithmetic average should not, for special reasons, be used. In general, however, arithmetic average is suitable for most purposes.

Limitations of Average.

In its very nature, an average is a summary of the details of a series ; it is a substitute for what it replaces. This is its limitation too. Different details may yield the same average, but the average will not reveal that the details were different. In fact, the details may turn out to be more significant than the average (see para 6, p. 116). Average does not relate the whole story about the series out of which it is computed. It does not show whether the differences of individual items from it are big or small. It indicates only the central position of the group. What lies behind it is not its task to reveal. Before crossing a river, the mathematician in the legend had made sure that the average height of the members of his family was equal to, or even slightly higher than, the

average depth of the river. It was only when all his family got sunk in the river while attempting to cross it, that he discovered that he could ill-afford to go by the average alone or ignore the individual measurements.

It follows that in computing and using an average one should know the following things to guard himself against a fallacy of argument :

(1) The purpose of the average.
(2) The peculiarities of the data to be summarized.
(3) The characteristics of each average.
(4) A deep knowledge of the whole subject to which the given data relate in order to be certain that the average computed shall be significant and suitable.
(5) The extent to which data are homogeneous.

QUADRATIC MEAN

The quadratic mean, also called the root-mean-square, is obtained by taking the square root of the average squares of the items of a series. If the items of a given series are 1, 3, 4, 5 and 7, the quadratic mean of the series will be

$$\sqrt{(1+9+16+25+49)\div 5}=4{\cdot}47$$

The arithmetic average of the given series is 4.

Quadratic mean is seldom used. We will see more of it while dealing with standard deviation in Chapter X.

MOVING AVERAGE

Moving average is an application of the simple arithmetic average. It is a simple method of indicating the trend of the given values over a period of time. A necessary preliminary to the calculation of moving average is to determine the period for which the moving average will be taken. This period should be long enough to eliminate the effect of seasonal or recurring fluctuations in the given time series. If the decision falls on a five-yearly moving average, the values for the first five year are totalled and their simple arithmetic average is taken. This average is placed opposite the third year which is the central year of the five. Then, the next five years' values are added up and averaged and the average placed against the

fourth year. That is, the first year is dropped and the years considered are from the second to the sixth, of which the fourth year is the central year. Then, the values for the third year to the seventh year are averaged and the average placed against the fifth year ; and, so on for the entire series (see col. 3, table 51).

In a seven-yearly moving average, the first average will be in respect of the first seven years and the average will be placed opposite the fourth year ; the second average will be for the second year to the eighth year and placed against the fifth year ; and so on. If the number of years for which the moving average is calculated is even, the average may be placed between the two middle years ; or, two adjacent moving averages may be added up and divided by 2 to arrive at the moving average for the middle year (see col. 5, table 51).

By calculating the moving average of a time series, a new series is obtained. This new series shows the trend, the cyclical fluctuations having been smoothened in the process of calculating the moving average. This trend may even enable a forecast for the near future and may thus provide a businessman with reasonably reliable basis for planning his policy for the immediate future. We will see more of this average in the Chapter on Analysis of Time Series.

PROGRESSIVE AVERAGE

This is another application of the method of simple arithmetic average. It differs from the moving average in that it is a cumulative average. When the average for, say, the first two years has been taken, the next average to be calculated will be for the first three years ; *i.e.,* the first year will not be dropped as is the case with the calculation of the moving average. Thereafter, the next average will be taken for the first four years ; and so on.

The progressive average is generally used during the early years of the working of a business, or until the business has run for a sufficiently long period to enable the calculation of moving averages. The profits, for instance, of each successive period may be compared with the progressive average profits for the entire previous period to determine how the business is progressing.

STANDARDIZED DEATH-RATE

If death-rate is calculated for each age-group of a locality's population, and then death-rate is calculated for the whole of the population by the use of weighted average, the latter death-rate is called **General Death-Rate** or **Crude Death-Rate.** If this crude death-rate for a locality is compared with that for the standard population (*e.g.*, the population of the country at large, or, of another locality assumed to be standard), misleading conclusions might result. To avoid fallacious comparison it is advisable to eliminate differences between age compositions of the populations of the two localities by applying the local death-rates in each age-group to the standard population. The following is a simple illustration—

Example 19. Required to compute crude and corrected death-rates.

Table 19. *Computation of General and Standardized Death-rates.*

Age-group Years	Standard Population A			Local Population B		
	Population	Deaths	Death-rate Per 1000	Population	Deaths	Death-rate Per 1000
Under 5	600	18	30	400	16	40
5—15	1000	5	5	1500	6	4
15—65	3000	24	8	2400	24	10
Above 65	400	20	50	700	21	30
Total	5000	67	13·4	5000	67	13·4

General Death-Rate of Standard Population

$$= \frac{1}{5000}(600\times30+1000\times5+3000\times8+400\times50)$$

$= 13{\cdot}4$ per 1,000.

General Death-Rate of Local Population

$$= \frac{1}{5000}(400\times40+1500\times4+2400\times10+700\times30)$$

$= 13{\cdot}4$ per 1,000.

Upon comparison of the two general death-rates, computed by using the weighted average, nothing remarkable will be noted: both the populations have the same death-rate. And, if death-rate is any measure of the health of a population, both the populations are equally healthy. To justify this viewpoint one might add that the total number of inhabitants in both the places, 5000, is the same, that the total number of deaths in the two cases, 67, is identical, and that the number of deaths in the age-group (15—65) years is equal in both of them. With these arguments one might try to make others believe that both A and B are equally healthy. But it should be noted that the death-rates in different age-groups in both the places are different and also that the distribution of population in the two places into various age-groups is not identical. That is, the basis of comparison is not the same. It is, therefore, not fully correct to believe that both the towns are equally healthy unless this conclusion is found to hold good when the basis of comparison is made identical. To do so, we eliminate the differences between age compositions by assuming that the distribution of local population into different age-groups is the same as that of the standard population. Then, by applying the local death-rates to the changed distribution we calculate another weighted average death-rate, now called the **Corrected** or **Standardized Death-Rate.** Thus,

Standardized Death-Rate of Local Population

$$= \frac{1}{5000}(600\times40+1000\times4+3000\times10+400\times30)$$

$= 14$ per thousand.

The standardized death-rate of local population is higher than the crude death-rate of standard population, leading us to conclude that the local population is less healthy than the standard population.

Similarly, there could be a case where the general death-rates of A and B were different, but the death-rates for different age constitutions were the same. This paradox, again, would have been due to differences in the distribution of population into the various age-groups. The paradox would be removed by computing the standardized or corrected death-rate of the local population.

This method is of general application. We have applied it to death-rates. We may standardize marriage rates or unemployment rates as well.

EXERCISES

(1) Define (*a*) arithmetic average, (*b*) geometric average, (*c*) median, and (*d*) mode. Which of these is the most representative, and why ?

(M. Com., Agra, 1945).

(2) What do you understand by homogeneity of data ? Should the data from which averages are computed be homogeneous ? Give reasons.

(3) What is a statistical average ? What are the desirable properties for an average to possess ? Which of the averages, you know, possesses most of these properties ?

(B. Com., Alld., 1944).
(M.A., Punjab, Apr. 1951).
(M.A., Delhi, 1953).

(4) Define weighted average and explain how it differs from simple mean. Give the method of its computation and point out the cases in which weighted average should be used.

(5) Differentiate between Crude (General) and Corrected (Standard) death-rates.

How is the principle of weighting applied to the determination of standardized death-rates from crude death-rates ?

(6) Discuss critically the use of weighted mean in statistics.

(B. Com., Cal., 1937).

(7) Explain clearly the meaning, the best method of computing, and the use of each of the following :—

(*i*) the arithmetic average.
(*ii*) the median, and
(*iii*) the mode.

(B. Com., Agra, 1934).

(8) What is the purpose served by an average ? What are the limitations of the uses of each one of the different kinds of average you know ?

(B. Com., Alld., 1948).

(9) What is the purpose served by an 'average'? Discuss the special advantages attached to the different averages, and illustrate their uses.

(B. Com., Agra, 1942).

(10) According to Census of India 1951 following are the population figures of cities in Part A States of India :—

(In thousand persons).

282	219	134	115	107	2,840
788	485	266	223	211	137
130	449	208	1,430	362	219
202	195	161	158	124	117
108	107	105	101	100	102
320	169	149	705	498	357
378	333	241	209	164	148
145	137	134	133	128	106
105	2,550	443	151	134	110
103					

Find the median and quartiles.

(11) Find the geometric and harmonic means of the series given in exercise 10, above.

(12) Compute the mode and median of the following series. Account for their difference.

Size of item	Frequency
2	3
3	8
4	10
5	12
6	16
7	14
8	10
9	8
10	17
11	5
12	4
13	1

(B. Com., Nagpur, 1945).

(13) Below are given the marks obtained by a batch of 20 students in a certain class-test in English and Hindi:—

Roll Number of Students	Marks in English	Marks in Hindi
1	53	58
2	54	55
3	52	25
4	32	32
5	30	26
6	60	85
7	47	44
8	46	80
9	35	33
10	28	72
11	25	10
12	42	42
13	33	15
14	48	46
15	72	50
16	51	64
17	45	39
18	33	38
19	65	30
20	29	36

In which subject is the level of knowledge of the students higher?

(M.A., Punjab, Sept. 1951).

(14) State the special characteristics and uses of the different kinds of averages. The values of cheques cleared at clearing houses in India in the years 1941-42 and 1942-43 were as follows:—

	Cheques cleared in 1941-42 (crores rupees)	Cheques cleared in 1942-43 (crores rupees)
April	181·5	205·4
May	219·6	213·0
June	188·2	182·8
July	193·0	235·3
August	244·6	181·3
September	213·0	187·4
October	236·6	261·1
November	233·0	238·7
December	228·6	261·6
January	262·7	323·9
February	221·2	282·4
March	213·8	339·9

Compute the different averages for both years and compare them. What comments have you to offer on the observed difference between the average monthly cheque clearances in the two years?

(M.A., Agra, 1945).

(15) Under what assumptions is mode located in a frequency distribution?

Compute the mode of the following series:—

Size of item	Frequency
4- 8	10
8-12	12
12-16	16
16-20	14
20-24	10
24-28	8
28-32	17
32-36	5
36-40	4

(B. Com., Alld., 1947).

(16) Find out the Mode of the following series:—

Size	Frequency	Size	Frequency
5	48	13	52
6	52	14	41
7	56	15	57
8	60	16	63
9	63	17	52
10	57	18	48
11	55	19	40
12	50	..	..

(B. Com., Alld., 1943).

(17) Compute the weighted geometric average of Relative Prices of the following commodities for the year 1939 (Base year: 1938 Price=100):—

Commodity	Relative Price	Weight (value produced in 1938)
Corn	128·8	1,385
Cotton	62·4	819
Hay	117·7	842
Wheat	99·0	561
Oats	130·9	408
Potatoes	143·5	194
Sugar	125·6	142
Barley	150·2	100
Tobacco	101·1	103
Rye	116·2	25
Rice	117·5	17
Oil Seed	78·7	29

How does it differ from the unweighted geometric mean and why?

(B. Com., Alld., 1943).

(18) 'Statistics help collective agreements of wage adjustments'. What data are required for the consideration of a revision in wage rates in a factory, which average will you utilize, and why?

(M.A., Agra, 1940).
(M. Com., Alld., 1943).

(19) Compare the relative advantages and disadvantages of the Arithmetic mean, the Median, and the Mode.

The following table gives the results of certain examinations of three universities in the year 1936. Which is the best university? Give reasons for your answer.

University Examination	Percentage results in the University		
	A	B	C
M.A.	80	75	70
M.Sc.	70	70	60
B.A.	65	80	70
B.Sc.	60	70	80
B.Com.	75	65	75

(M.A., Cal., 1937).

(20) The following table gives the heights of 1,000 students:—

Height (Centimeters)	Number of students	Height (Centimeters)	Number of students
155—157	4	179—181	125
158—160	8	182—184	92
161—163	26	185—187	60
164—166	53	188—190	22
167—169	89	191—193	4
170—172	146	194—196	1
173—175	188	197—199	1
176—178	181	..	..
		Total ..	1,000

(*a*) Compute the value of the median, quartiles, and third decile.

(*b*) Draw an ogive and verify your results.

(M.A., Alld., 1950).

(21) Draw a cumulative frequency graph of the following distribution showing the monthly wages of a group of workmen and hence or otherwise calculate the values of (*a*) the mode, (*b*) the median, and (*c*) the two quartiles:—

Wages in Rs.		21-	22-	23-	24-	25-	26-	27-	28-29
No. of Workmen	8	10	11	16	20	25	15	9	6

(M.A., Raj., 1950).

(22) The following table gives the number of persons with different incomes in the U.S.A. during the year 1929.

Income in thousands of dollars.	No. of persons in Lakhs.
Under 1	13
1 — 2	90
2 — 3	81
3 — 5	117
5 — 10	66
10 — 25	27
25 — 50	6
50 — 100	2
100 —1000	2

Calculate the average income per head.

(B. Com., Luck., 1939).

(23) Explain what is meant by *weighted average*, and discuss the effect of weighting.

Calculate (i) the *unweighted mean* of the prices in column III and (ii) the mean obtained by weighting each price by the quantity consumed, and explain why they differ as they do:—

I Articles of Food	II Quantity Consumed	III Price in Rupees per maund
Flour	11·5 mds.	5·8
Ghee	5·6 mds.	58·4
Sugar	·28 mds.	8·2
Potato	·16 mds.	2·5
Oil	·35 mds.	20·0

(M.A., Cal., 1937).

(24) Explain the short-cut method of calculating the arithmetic average.

The following data relate to sizes of shoes sold at a store during a given week. Find the average size by the short-cut method:

Size of Shoes	4·5	5	5·5	6	6·5	7	7·5	8	8·5	9	9·5	10	10·5	14
No. of Pairs	1	2	4	5	15	30	60	95	82	75	44	25	15	1

(M.A., Cal., 1936).

(25) From the following table calculate the average price of a lb. of biscuits and also the weighted average price.

Price per lb. Rs. a. p.	lbs. sold
0—10—0	100
0—15—0	87
1— 2—0	63
1— 4—0	59
1— 8—0	49
2— 0—0	19

Which of the two averages gives a better indication of the average price ?

(26) Compute the geometric and harmonic means of the following items.

375.5, 15.3, 28.5, 12.01, 4.5, 3.7, 12.79, 35, 41.9, 53.

How will the weighted geometric and harmonic means compare if the weights were, respectively, 2, 10, 8, 5, 9, 7, 4, 6, 3 and 1 ?

(27) If in ten successive years the quantities 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 are sold at prices 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, what are the weighted average and simple arithmetic average prices?

(28) Show the relative positions of different averages in moderately symmetrical series.

Find the mode, median and arithmetic average of the following series and state if the series is symmetrical.

Size of item	Frequency
10	30
11	35
12	38
13	42
14	46
15	42
16	38
17	35
18	30

(29) Draw an ogive showing the number of candidates obtaining marks higher than any assigned value from the following data:—

Marks obtained	0—10	11—20	21—30	31—40	41—50	51—60	61—70	71—80	81—90	91—100
No. of Candidates	33	108	221	439	526	495	322	153	53	7

Read off the median and the two quartiles from the curve and check your results by actual calculation. Compare the median with the mean and comment on the difference if any.

(B. Com. Hons., Andhra, 1943).

(30) Find the Median, Lower Quartile and 67th Percentile of the frequency distribution given below:—

Marks in Statistics

Marks-Group	Number of Students
Under 10	8
10—20	12
20—30	20
30—40	32
40—50	30
50—60	28
60—70	12
70 and above	4

(B. Com., Alld., 1949).

(31) The following are the weekly market values of the shares of 'Imperial Bank of India' (paid up value Rs. 500) from Jan. 4, 1933 to Dec. 30, 1933.

1235, 1235, 1236.5, 1261.5, 1266.5, 1166, 1190, 1176, 1160, 1186, 1221.5, 1220, 1234, 1230, 1235, 1236.5, 1232.5, 1251.5, 1244, 1231.5, 1216.5, 1216.5, 1221.5, 1221.5, 1205, 1184, 1196, 1212.5, 1207, 1192, 1195, 1206, 1201, 1196, 1198, 1208.5, 1202.5, 1234.5, 1234.5, 1225, 1230, 1220, 1220.5, 1220.5, 1240, 1252.5, 1246, 1246.5, 1246.5, 1230.5.

Calculate the arithmetic mean and geometric average of the above series.

(32) The following table gives the age distribution of widows in undivided India, (Census Report 1931). Calculate the median age of the widows and also the upper and lower quartiles.

Years			No. of widows
0—10	..	..	135,862
10—20	..	..	718,101
20—30	..	..	2,456,835
30—40	..	..	4,847,631
40—50	..	..	6,480,259
50—60	..	..	5,908,159
60—70	..	..	3,743,615
70 & over	..	..	1,957,506
	Total	..	26,247,968

(33) The following table gives the marks obtained by a batch of 30 B. Com. students in a class-test in Statistics. (Marks 100).

Roll No.	Marks obtained	Roll No.	Marks obtained
1	33	16	24
2	32	17	33
3	55	18	42
4	47	19	38
5	21	20	45
6	50	21	26
7	27	22	33
8	12	23	44
9	68	24	48
10	49	25	52
11	40	26	30
12	17	27	58
13	44	28	37
14	48	29	38
15	62	30	35

Find the values of the Mode, the Median, and the Quartiles.
(B. Com., Alld., 1938).

(34) Calculate the arithmetic average and the geometric mean of the series given in exercise 33 above. Which average represents the series better ?
(B. Com., Alld., 1938).

(35) The following table gives the distribution of the male population of a certain area in India. Find the mean age and median age.

Age Group	Males
0— 9	2,756
10—19	2,124
20—29	1,677
30—39	1,481
40—49	1,021
50—59	610
60—69	245
70—79	67
80—89	16
90—99	3
	10,000

(B. Com., Alld., 1951).

(36) The following table gives the male population of Kanpur and Jaipur in 1931:—

Age-group (years)	Population of males in thousands	
	Kanpur	Jaipur
0— 5	14	9
5—10	13	8
10—15	13	8
15—20	13	7
20—30	33	15
30—40	29	12
40—50	17	9
50—60	7	6
60—80	4	4

Calculate the average age of males at Kanpur and Jaipur separately, and account for the difference, if any.

(B. Com., Alld., 1952).

(37) The following table gives the distribution of population according to age in India (undivided) and Japan at the time of the last census (1931):

Age-group in years	Population in millions in India	Population in millions in Japan
0—10	98·9	17·8
10—20	72·5	14·3
20—30	63·2	11·3
30—40	48·6	8·6
40—50	32·6	6·5
50—60	19·4	5·4
60—80	13·2	5·1

Calculate the average age of people in India and in Japan, and comment on the difference.

(B. Com., Alld., 1940).

(38) Calculate the simple average and the weighted average of the following items:—

Item	68	85	101	102	108	110	112	113	124	128	143	146	151	153	172
Weight	1	46	31	1	11	7	23	17	9	14	2	4	6	5	2

Account for the difference in the two averages.

(M.A., Alld., 1940).

(39) The following is the distribution of wages per thousand employees in a certain factory:

Daily wages in Annas	2-	4-	6-	8-	10-	12-	14-	16-	18-	20-	22-	24-	Total
Number of employees	3	13	43	102	175	220	204	139	69	25	6	1	1000

Calculate the modal and median wages, and explain why there is a difference between the two.

(M.A., Alld., 1940).

(40) Calculate the average, median, and upper and lower quartile ages in the following table:

Age-groups	Population in thousands	
	1881	1931
0— 4	3520	3280
5— 9	3160	3500
10—19	5340	7200
20—29	4560	6640
30—39	3420	5980
40—49	2660	5240
50—59	1900	3780
60—69	1320	2440
70—79	600	1220
80 & over	120	320

(M.A., Agra, 1940).

(41) Point out the ambiguity or mistake, if any, in the following statements:

(*i*) There are 250 employees in a sugar mill. Their daily earnings are about Re. 1/- per man on an average. Therefore, their total monthly earnings are Rs. 7,500.

(*ii*) An ordinary person in India consumes one chhatak of pulse per day. Therefore, the total quantity of pulse consumed by India's 40 crores of people every year is about 23 crores of maunds.

(*iii*) The monthly expenditure of the vast majority of students in a university is Rs. 50. Therefore, the total monthly expenditure of 2,000 students is Rs. 1,00,000.

(*iv*) A cloth dealer usually receives 150 customers a day. Therefore, the total number of customers he receives in a month is 4,500.

(42) Find out the average height of a clerk in a certain office from the following figures. What is the median height and how far does it differ from the mode?

Height	Frequency	Height	Frequency
5′ 6″	1	5′ 11″	1
5′ 7″	2	6′ 0″	2
5′ 8″	4	6′ 1″	1
5′ 9″	3	6′ 2″	1
5′ 10″	2	6′ 3″	1

(M.A., Agra, 1937).

(43) A railway train runs for 50 minutes at a speed of 40 miles an hour and then, because of repairs of the track, runs for 10 minutes at a speed of 10 miles an hour. What is its average speed?

(44) The following is the age distribution of candidates appearing at the Matriculation and Intermediate Arts examinations of the Patna University in 1937.

Age in years	12-	13-	14-	15-	16-	17-	18-	19-	20-	21-	22-	Total
Matriculation	5	48	189	303	522	980	981	794	515	474	×	4811
Intermediate	×	×	×	5	45	87	127	150	155	127	175	871

Compare the median and modal ages of the Matriculation candidates with those of I.A. candidates.

(M.A., Patna, 1940).

(B. Com., Alld., 1947).

(45) Amend the following table, and locate the median from the amended table. Also measure the magnitude of the median so located.

Sizes	Frequency
10—15	10
15—17·5	15
17·5—20	17
22—30	25
30—35	28
35—40	30
45 & upwards	40

(B. Com., Alld., 1942).

(46) Monthly incomes of twenty families are given below in rupees:—

2,000; 35; 400; 15; 40; 1,500; 300; 6; 90; 250; 20; 12; 450; 10; 150; 8; 25; 30; 1,200; 60.

Calculate the Geometric Mean and the Harmonic Mean of the above incomes.

(B. Com., Alld., 1941).

(47) What considerations will weigh with you in choosing a suitable average for studying a phenomenon? Give a few typical cases in which your choice will fall on averages other than the arithmetic average.

(48) An average is a substitute for a complex group of variables; but it is not always safe to depend on the substitute alone, to the exclusion of individual measurements of the group.

Explain and illustrate the above statement.

(49) (*a*) If the mode and mean of a moderately asymmetrical series are, respectively, 18 inches and 17·7 inches, what would be its most probable median?

(*b*) A group of 25 students earned an average of 35 marks in a certain test. A second group of 75 students averaged 45 marks in the same test. What was the average score of the 100 students?

(*c*) If in a series which is not highly skewed the mean deviation is 9·8 feet, what would be the approximate value of its standard deviation?

(B. Com., Alld., 1946).

(50) What are the types in statistical investigation? Discuss the relative merits of different types (kinds of average) used.

(B. Com., Luck., 1943).

(51) Enumerate and define the various kinds of averages in common use, and indicate the purpose for which they are respectively required. What fallacies have to be guarded against in using averages?

(M.A., Agra, 1930).

(52) The following is the distribution of wages per thousand employees in a certain factory:—

Daily wages in annas	No of employees
2—	3
4—	13
6—	43
8—	102
10—	175
12—	220
14—	204
16—	139
18—	69
20—	25
22—	6
24—	1

Calculate the modal and the median wages, and explain why there is a difference between the two.

(M.A., Alld., 1940).

(53) Define the following averages, and state briefly the advantages and disadvantages of each, namely, (*a*) mode, (*b*) median, (*c*) geometric average, (*d*) simple arithmetic average, (*e*) weighted arithmetic average.

In what circumstances are the simple and weighted arithmetic averages likely to be equal.

(M.A., Cal., 1936).

(54) In order to decide whether one city is more healthy than another on the basis of death-rates what information would you require in addition to the total number of deaths and the total population of the two cities? How would you use this information to decide which city is more healthy?

(M.A., Alld., 1935).

(55) Which of the two places for which the mortality data are given below would you describe as more healthy than the other? Give reasons.

Age in Years	Town X (Standard)		Town Y (Local)	
	Population	Deaths	Population	Deaths
under 10	15,000	375	10,000	300
10—30	50,000	250	52,000	312
30—70	120,000	840	126,000	1,008
above 70	15,000	975	12,000	840

(56) Compute the Crude and Standardized death-rates in the following, and state if local population has higher or lower death-rate.

Age group (years)	Standard Population		Local Population	
	Population	Deaths	Population	Deaths
under 5	6,000	160	2,500	63
5—15	10,000	20	12,500	25
15—65	12,500	50	20,000	80
above 65	4,000	160	5,000	200

(57) The following are the death-rates per thousand per annum of two towns in a certain year:

Ages (years)	Town A			Town B		
	Population	Deaths	Death Rate per 1,000	Population	Deaths	Death Rate per 1,000
under 2	3,000	192	64·0	5,000	300	60·0
2—10	10,000	70	7·0	12,000	78	6·5
10—20	10,000	40	4·0	10,000	38	3·8
20—60	32,500	260	8·0	25,000	190	7·6
60 & above	8,500	510	60·0	8,000	460	57·5
All	64,000	1,072	16·75	60,000	1,066	17·77

(*a*) For each age-group the death-rate of Town A is greater than that of Town B, but the reverse is the case when all age-groups are grouped together. Why is it so ?

(*b*) Calculate the standardized death rate for Town B, taking the population of Town A as the standard.

(B. Com., Hons., Andhra, 1944).
(B. Com., Alld., 1946).
(M.A., Pb., Sept., 1951).

(58) Write brief explanatory notes on the following :—

(*a*) Averaging is the process of condensation ;
(*b*) Moving Average ;
(*c*) Progressive Average ;
(*d*) Quadratic Mean ;
(*e*) Typical and Descriptive averages.

(59) Compare the merits and demerits of the median and the mode. In which of the following problems would they be most useful ?—

(*a*) Skull measurements ;
(*b*) Size of holdings ;
(*c*) Comparison of intelligence ;
(*d*) Marks obtained in any examination ;
(*e*) Heights and weights of students.

(M.A., Agra, 1943).

CHAPTER X

DISPERSION AND SKEWNESS

DISPERSION

Averages of Second Order.

Averages of the "first order", discussed in the last Chapter, consider only the central position of a series. They do not throw light on the formation of the series. They fail to characterize the details from which they are made. Hardly ever are the various items of a series equal to the value of the average computed from them. Some measure of the differences of the items from their average is necessary. Averages of the "second order" provide this measure. An average of the second order is an average of the differences of all the items of the series from an average of those items. In averaging these differences their irregularities are brushed off and a representative figure results.

All frequency distributions are not similar. They may differ in two ways. Firstly, they may differ in the numerical size of their averages though not necessarily in their formations. For instance, the averages of two series—1, 2, 3 and 7, 8, 9—are respectively 2 and 8, and therefore different; but the scatteredness of the items in the two series around their respective means is similar: the greatest deviation from the average in either series is only 1. The two series are, therefore, not differently formed, and the difference in their averages can be studied with the help of averages of the 'first order'.

Secondly, frequency distributions may have the same values of their averages yet differ in their respective formations. Let us suppose that the daily earnings of A and B, two mechanics, during the six days of a working week are as given below :

Days	A's earnings Rs.	B's earnings Rs.
Monday ..	4	3
Tuesday ..	4	4
Wednesday ..	5	5
Thursday ..	5	5
Friday ..	6	6
Saturday ..	6	7
6 days	Rs. 30	Rs. 30

The total earnings made in 6 days, Rs. 30, and therefore, the average earnings, Rs. 5 per day, are thus exactly the same in both the cases, but the scatteredness of the values of the items of the series round their average earnings is different in the two cases. The greatest deviation from the average in A's case is Re. 1 and in B's case Rs. 2. The two series are, therefore, differently constituted, though they result in averages of the same numerical value. The difference in the nature of formation of the two series can be studied with the help of averages of the 'second order'.

The above example of daily earnings of A and B provides an illustration of how averages of the 'first order' may prove to be misleading and how averages of the 'second order' keep us from being misled. The fact that the variability is greater in the case of B discounts the significance of the two averages being identical: On an average both A and B earn the same amount, but A is more consistent than B; B's earnings show comparatively greater fluctuations inasmuch as sometimes he earns as low as Rs. 3 and sometimes as high as Rs. 7 per day.

It follows from the above that averages must be used with great caution. To these cautions belongs the measurement of the dispersion or scatteredness of the series around the mean. The value of the mean depends necessarily on the distribution of the items around it and on its position in the series. An examination of this distribution furnishes us with a valuable

supplement to the information given by the mean itself: it tells us how the items comprised in a group vary in size. This helps us in finding the extent to which the average is 'typical', for, the less the variability of a series the more typical is its average.

Meaning of Dispersion.

The term "Dispersion" is used in two senses in Statistics. One sense is general, implying that within a given group the items are not uniform in their size. That is, they differ in their magnitude. This difference may be great or small. Accordingly, dispersion may be considerable or slight. If the profits of a given number of businesses in the same trade and with the same capital are found to vary between Rs. 11,000 and Rs. 11,020, they are scattered over a small range. That is, they are fairly consistent, or, in other words, their variability is slight. If, however, the profits vary between Rs. 1,100 and Rs. 11,000, consistency is wanting, the range is wide, or, in other words, variability is considerable.

The other sense in which the term dispersion is used is more precise. In this sense it indicates an absolute or relative measure of the differences of the items of a group from some average computed from those items. It may be noted that the difference between the measurements of the value of a variable and its mean, or any other fixed point, is technically termed deviation. And, a measure of the deviations of the size of items from an average is called the **Measure of Dispersion.**

Measures of Dispersion.

The two senses in which the term dispersion is used are important. The first sense points to the limits within which data fall; the second sense calls attention to the amount, absolute or relative, by which the values of the items differ from an average or type. The two senses are different. Dispersion, in the first sense, is indicated by the method of limits, where the complete range of the items may be shown. Dispersion, in the second sense, is expressed by the method of averaging differences from a type.

METHOD OF LIMITS.

The most common way of measuring dispersion under this method is that of computing the Range.

The Range.

Range of dispersion represents the difference between the values of the extremes, *i.e.,* the largest and the smallest items, of the data under review. If in a certain class the height of the shortest student were 4 ft. 10 in., and that of the tallest 6 ft., the range would evidently be 14 inches. Range is thus the simplest method of measuring dispersion; but it is too indefinite to be used as a practical measure of dispersion, since it depends entirely upon the values of the extreme items. For instance, if a dwarf whose height was only 3 ft. 6 inches were admitted to the class, in the above example, the range would suddenly rise to 30 inches, while the average height of the class would not be materially affected. There is yet another reason why the range is not a satisfactory measure of dispersion. It is that the range does not take into account the distribution of the items within its limits. This distribution may vary widely, even though the extremes be of the same value. We might get the same value for the range from a symmetrical and a J-shaped (*i.e.,* asymmetrical) frequency-curve. Clearly, two such distributions cannot be regarded to exhibit the same dispersion. On the other hand, series with different maxima and minima may have practically the same distribution or dispersion.

Besides, the range is an absolute measure of dispersion, and, therefore, it is not possible to compare through it the *relative* dispersion of two series expressed in different units. Absolute dispersion measured, say, in yards is not comparable with that measured in tons. If the absolute dispersions are reduced to relative bases, comparison would be possible. This is done by dividing the range by the sum of the extreme items. The quotient is called the **co-efficient or ratio of dispersion.**

In our example of the height of students taken above, the range is 14 inches, and the co-efficient is $\frac{14}{130} = 0{\cdot}108$.

METHOD OF AVERAGING DEVIATIONS.

Under this method, the most common measures are :

(1) The Average or Mean Deviation, also called the First Moment of Dispersion;

(2) The Standard Deviation; and

(3) The Quartile Deviation.

These are absolute measures. They are converted into relative measures, called coefficients, for purposes of rendering comparison possible between series measured in different units.

Mean Deviation and its Co-efficient.

The first moment of dispersion, also called the mean or average deviation, is the arithmetic average of the deviations of the group measured from an average (Median, Mode or Mean) taking all deviations as positive. In other words, mean **deviation is the sum of the deviations from an average divided by the number of items.** It is necessary to treat all deviations as positive, since the sum of the deviations from the arithmetic average taken with minus and plus signs is zero, and that from the median and the mode is nearly zero.

Let d stand for deviation, *i.e.*, difference between an individual item and the average, and m for the value of an individual item. Then,

$d = m - a =$ deviation from the arithmetic average,

$d_m = m - M =$ deviation from the median,

$d_z = m - Z =$ deviation from the mode.

If n be the number of items in a series,

$\frac{\Sigma d}{n} =$ The First Moment of Dispersion from the Arithmetic Average, represented by δ (Delta).

$\frac{\Sigma d_m}{n}$ = The First Moment of Dispersion from the Median, represented by δ_m.

$\frac{\Sigma d_z}{n}$ = The First Moment of Dispersion from the Mode, represented by δ_z.

It is proper to calculate the mean deviation from the median because the sum of deviations, and consequently the mean deviation, is *least* when median is chosen as the origin from which deviations are measured. In practice, however, it is easier and not un-satisfactory to calculate it from the arithmetic average. In case the observations are recorded in groups between different limits, mean deviation from the median is difficult to calculate with precision, and in that case, arithmetic mean rather than the median may be chosen as the origin. It is not common practice to calculate the mean deviation from the mode.

Mean deviation gives us the absolute measure of dispersion. This is one factor that is required for calculating the relative measure of dispersion, called **Mean Co-efficient of Dispersion.** The other factor required is the mean used in the particular case. Thus,

$\frac{\delta}{a}$ = Co-efficient of Dispersion from the Arithmetic Average,

$\frac{\delta_m}{M}$ = Co-efficient of Dispersion from the Median, and

$\frac{\delta_z}{Z}$ = Co-efficient of Dispersion from the Mode.

Calculation of Mean Deviation and its Co-efficient.

Example 1. Required to find mean deviation and its co-efficient when individual quantitative observations are given.

Table 20. *Calculation of Mean Deviation of X's Monthly Earnings for a year.*

Months	Monthly Earnings m	Deviations from median (Rs. 42) (+&−signs ignored) d_m
	Rs.	Rs.
1	39	3
2	40	2
3	40	2
4	41	1
5	41	1
6	42	0
7	42	0
8	43	1
9	43	1
10	44	2
11	44	2
12	45	3
$n=12$		$\Sigma d_m=$Rs. 18

Median or $M=$ Value of $\left(\frac{n+1}{2}\right)^{\text{th}}$ item

$=$ Value of 6·5th item

$=$ Rs. 42.

Mean Deviation from the Median, or, $\delta_m=\frac{\Sigma d_m}{n}$

$=$ Rs. $\frac{18}{12}$

$=$ Rs. 1·5

Co-efficient of Dispersion from the Median $=\frac{\delta_m}{M}=\frac{1·5}{42}$

$=$ ·0357 approximately.

Since the arithmetic mean and the median have the same values in this example, mean deviation from the arithmetic mean and its co-efficient of dispersion will also have the same

values as those from the median. We can now state that the mean or median earnings of X are Rs. 42 and the earnings deviate from the mean or median on an average by Rs. 1·5.

Example 2. Required to find mean deviation and its co-efficient when a discrete series is given.

Table 21. *Calculation of Mean Deviation.*

Size of item m	Frequency f	Deviation from Median (5) (+&−signs ignored) d_m	Total Deviation (Frequency× deviation) fd_m
2	2	3	6
3	3	2	6
4	5	1	5
5	8	0	0
6	6	1	6
7	4	2	8
8	2	3	6
9	1	4	4
	$n=31$		$\Sigma d_m=41$

Median or M = the value of $\left(\frac{n+1}{2}\right)^{\text{th}}$ item,

= the value of 16th item,

= 5

Mean Deviation from Median or $\delta_m = \frac{\Sigma d_m}{n}$

$$= \frac{41}{31} = 1{\cdot}32$$

Mean Co-efficient of Dispersion $= \frac{\delta_m}{M}$

$$= \frac{1{\cdot}32}{5} = {\cdot}264$$

Example 3. Required to find the mean deviation and its co-efficient when the data are composed of a continuous series, measuring the deviations from the median as well as from the arithmetic average.

Table 22. *Calculation of Mean Deviation of Marks of* 60 *Students in Economics.*

Marks-group	Mid-value m	Frequency f	Deviation from Median (35 marks) (+&—signs ignored) d_m	Total Deviation from Median fd_m	Deviation from Arithmetic Average (34·7 marks) (+&—signs ignored) d	Total Deviation from Arithmetic Average fd
0—10	5	4	30	120	29·7	118·8
10—20	15	8	20	160	19·7	157·6
20—30	25	11	10	110	9·7	106·7
30—40	35	15	0	0	0·3	4·5
40—50	45	11	10	110	10·3	113·3
50—60	55	7	20	140	20·3	142·1
60—70	65	4	30	120	30·3	121·2
		$n=60$		$\Sigma d_m=760$		$\Sigma d=764·2$

(*i*) Median or $M=35$ marks, by interpolation in (30-40) marks group.

Mean Deviation from Median or $\delta_m=\frac{\Sigma d_m}{n}=\frac{760}{60}$ marks

$=12·67$ marks, approx.

Co-efficient of Dispersion from Median$=\frac{\delta_m}{M}=\frac{12·67}{35}$

$=·36$ approx.

(*ii*) Arithmetic Average or $a=34·7$ marks.

Mean Deviation from Arithmetic Average or $\delta=\frac{\Sigma d}{n}$

$=\frac{764·2}{60}$ marks

$=12·74$ marks approx.

$$\text{Co-efficient of Dispersion from Arithmetic Average} = \frac{\delta}{a} = \frac{12{\cdot}74}{34{\cdot}7} = {\cdot}37 \text{ approx.}$$

The above example amply demonstrates that mean deviation when measured from the median is *least*, for δ_{mz} is less than δ and it would also be less than δ_z, or mean deviation from any other point.

Characteristics and Uses of Mean Deviation and its Co-efficient.

Mean deviation and mean co-efficient of dispersion are easy to calculate and comprehend; they take all items into consideration and give weight to deviations according to their size. The co-efficient is usefully employed in economic studies like the distribution of personal wealth in a community where the rich and the poor both are considered. But the mean deviation does not lend itself readily to algebraical treatment. Other moments of dispersion have, therefore, come into use.

Standard Deviation and its Co-efficient.

An alternative method of eliminating the algebraical signs of the deviations from an average is to square up each deviation. This method is employed here. **Second moment of dispersion** is the sum of the squares of the individual deviation from the arithmetic average divided by their number, *viz.*, $\frac{\Sigma d^2}{n}$. **Standard deviation is the square root of the Second Moment**, *viz.*, $\sqrt{\text{Second Moment}}$ or $\sqrt{\frac{\Sigma d^2}{n}}$ represented by σ (Sigma). Standard deviation is an absolute measure of dispersion and is *least* when arithmetic average is chosen as the origin from which deviations are measured. To compute the **Standard Co-efficient of Dispersion**, a relative measure, the standard deviation is divided by the arithmetic average, *viz.*, $\frac{\sigma}{a}$

Similar moments and absolute measures of dispersion based on mode and median are quite conceivable. Second moment of dispersion computed from any mean or value other than the arithmetic average is sometimes termed **Mean Square Deviation** and the absolute measure of dispersion based on such second moment is designated **Root-Mean Square Deviation.** But root-mean square deviation is converted into standard deviation before any use is made of it. Since the sum of the squares of the deviations from the arithmetic average is minimum, it is obvious that standard deviation is that root-mean square deviation whose value is the least, and second moment is that mean square deviation whose value is minimum.

Calculation of Standard Deviation and its Co-efficient.

Direct Method

Example 4. Required to find standard deviation and its co-efficient when quantitative individual observations are given.

Table 23. *Calculation of Standard Deviation of X's Monthly Earnings for 12 months by the Direct Method.*

Months	Monthly earnings m	Deviation from Arithmetic Average (Rs. 42) d	Square of Deviation d^2
1	39	−3	9
2	40	−2	4
3	40	−2	4
4	41	−1	1
5	41	−1	1
6	42	0	0
7	42	0	0
8	43	+1	1
9	43	+1	1
10	44	+2	4
11	44	+2	4
12	45	+3	9
$n=12$	$\Sigma m=$Rs. 504 $a=$Rs. 42		$\Sigma d^2=38$

Arithmetic Average or $a = \frac{\Sigma m}{n} = \text{Rs. } \frac{504}{12} = \text{Rs. } 42$

Second Moment of Dispersion $= \frac{\Sigma d^2}{n} = \frac{38}{12}$
$= 3{\cdot}17$ Sq. Rs. approx.

Standard Deviation or $\sigma = \sqrt{\frac{\Sigma d^2}{n}} = \sqrt{3{\cdot}17} = \text{Rs. } 1{\cdot}78$ approx.

Standard Co-efficient of Deviation $= \frac{\sigma}{a} = \frac{1{\cdot}78}{42} = {\cdot}042$ approx.

We can now state that the arithmetic average of the given series is Rs. 42 and its standard deviation is Rs. 1·78.

Example 5. Required to find standard deviation and its co-efficient when a discrete series is given.

Table 24. *Calculation of Standard Deviation by the Direct Method.*

Size of items m	Frequency f	Sum of sizes (Frequency × size of item) mf	Deviation from mean (8) d	Square of deviation d^2	Product of square of deviation and frequency fd^2
2	1	2	−6	36	36
4	2	8	−4	16	32
6	3	18	−2	4	12
8	5	40	0	0	0
10	3	30	+2	4	12
12	2	24	+4	16	32
14	1	14	+6	36	36
	$n=17$	$\Sigma m=136$ $a=8$			$\Sigma d^2=160$

Arithmetic Average or $a = \frac{\Sigma m}{n} = \frac{136}{17} = 8$

$$\text{Second Moment of Dispersion} = \frac{\Sigma d^2}{n} = \frac{160}{17} = 9{\cdot}4 \text{ approx.}$$

$$\text{Standard Deviation or } \sigma = \sqrt{\frac{\Sigma d^2}{n}} = \sqrt{9{\cdot}4} = 3{\cdot}066 \text{ approx.}$$

$$\text{Standard Co-efficient of Deviation} = \frac{\sigma}{a} = \frac{3{\cdot}066}{8} = {\cdot}383 \text{ approx.}$$

We can state that the arithmetic average of the given series is 8, and its standard deviation is 3·066.

Example 6. Required to find the standard deviation and its co-efficient when a continuous series is given.

Suppose the class-intervals of a given series are 1-3, 3-5, 5-7, 7-9, 9-11, 11-13, 13-15 and their respective frequencies are 1, 2, 3, 5, 3, 2, 1. Then, the class-intervals will be placed in the first column of a table, their mid-points 2, 4, 6, 8, 10, 12, 14, in the second column, which tally with the size of items of table 24, example 5, above. The entire working of the example will thereafter be the same as that in example 5, resulting, in this particular case, in the same values of arithmetic average, second moment, standard deviation and standard co-efficient of dispersion.

Short-Cut Method.

The preceding direct method of computing the standard deviation is easy and simple if the arithmetic average of the series is a whole number, as it was in examples 4, 5 and 6, above. But, often the arithmetic average happens to contain a fraction. Then, the deviations from the true mean would also contain fraction. The calculation of such deviations and their squaring up will be tedious. In such cases the short-cut method of computing the standard deviation can be usefully employed in place of the direct method.

The short-cut method of computing the standard deviation, as we shall presently see, saves valuable time and tiresome effort. It gives the same value for the standard deviation as the direct method does. The procedure is as follows: Assume any size of item as the average; compute deviations from it;

square each deviation; summate; subtract from the summation n times the square of the difference between the assumed average and the true average; divide by n and extract the square root of the quotient. Symbolically,

$$\sigma = \sqrt{\frac{\Sigma d^2_x - n\ (a-x)^2}{n}} \qquad \text{or, } \sigma = \sqrt{\frac{\Sigma d^2_x}{n} - (a-x)^2}$$

where, x represents assumed average[1] ; a, actual average ; d^2 square of deviation from assumed average ; and n, number of observations.

Example 7. Required to calculate standard deviation of a continuous series. (The procedure worked out in this example shall also apply to series of any other type).

Table 25. *Calculation of Standard Deviation by the Short-Cut Method.*

(1) Size of item	(2) Mid-value m	(3) Frequency f	(4) Sum of sizes [col. (2) × col. (3)] mf	(5) Deviation from the assumed average (5) d_x	(6) Square of Dev. d^2_x	(7) Product of Square of Dev. & Frequency [col. (3) × col. (6)] fd^2_x
1.5—2.5	2	3	6	—3	9	27
2.5—3.5	3	4	12	—2	4	16
3.5—4.5	4	5	20	—1	1	5
4.5—5.5	5	8	40	0	0	0
5.5—6,5	6	7	42	+1	1	7
6.5—7.5	7	6	42	+2	4	24
7.5—8.5	8	3	24	+3	9	27
		n=36	Σm=186			Σd^2_x=106

[1] The assumed average x i.e. is a whole number approximating the actual average. That value which has the maximum frequency may also be taken as the assumed average, or the working mean as it is often called.

$$\text{Second Moment of Dispersion} = \frac{\Sigma\, d^2}{n} = \frac{160}{17} = 9{\cdot}4 \text{ approx.}$$

$$\text{Standard Deviation or } \sigma = \sqrt{\frac{\Sigma\, d^2}{n}} = \sqrt{9{\cdot}4} = 3{\cdot}066 \text{ approx.}$$

$$\text{Standard Co-efficient of Deviation} = \frac{\sigma}{a} = \frac{3{\cdot}066}{8} = {\cdot}383 \text{ approx.}$$

We can state that the arithmetic average of the given series is 8, and its standard deviation is 3·066.

Example 6. Required to find the standard deviation and its co-efficient when a continuous series is given.

Suppose the class-intervals of a given series are 1-3, 3-5, 5-7, 7-9, 9-11, 11-13, 13-15 and their respective frequencies are 1, 2, 3, 5, 3, 2, 1. Then, the class-intervals will be placed in the first column of a table, their mid-points 2, 4, 6, 8, 10, 12, 14, in the second column, which tally with the size of items of table 24, example 5, above. The entire working of the example will thereafter be the same as that in example 5, resulting, in this particular case, in the same values of arithmetic average, second moment, standard deviation and standard co-efficient of dispersion.

Short-Cut Method.

The preceding direct method of computing the standard deviation is easy and simple if the arithmetic average of the series is a whole number, as it was in examples 4, 5 and 6, above. But, often the arithmetic average happens to contain a fraction. Then, the deviations from the true mean would also contain fraction. The calculation of such deviations and their squaring up will be tedious. In such cases the short-cut method of computing the standard deviation can be usefully employed in place of the direct method.

The short-cut method of computing the standard deviation, as we shall presently see, saves valuable time and tiresome effort. It gives the same value for the standard deviation as the direct method does. The procedure is as follows: Assume any size of item as the average; compute deviations from it;

square each deviation; summate; subtract from the summation n times the square of the difference between the assumed average and the true average; divide by n and extract the square root of the quotient. Symbolically,

$$\sigma = \sqrt{\frac{\Sigma d^2_x - n\,(a-x)^2}{n}} \quad \text{or,} \quad \sigma = \sqrt{\frac{\Sigma d^2_x}{n} - (a-x)^2}$$

where, x represents assumed average[1] ; a, actual average ; d^2 square of deviation from assumed average ; and n, number of observations.

Example 7. Required to calculate standard deviation of a continuous series. (The procedure worked out in this example shall also apply to series of any other type).

Table 25. *Calculation of Standard Deviation by the Short-Cut Method.*

(1) Size of item	(2) Mid-value	(3) Fre-quency	(4) Sum of sizes [col. (2) × col. (3)]	(5) Deviation from the assumed average (5)	(6) Square of Dev.	(7) Product of Square of Dev. & Frequency [col. (3) × col. (6)]
	m	f	mf	d_x	d^2_x	fd^2_x
1.5—2.5	2	3	6	—3	9	27
2.5—3.5	3	4	12	—2	4	16
3.5—4.5	4	5	20	—1	1	5
4.5—5.5	5	8	40	0	0	0
5.5—6,5	6	7	42	+1	1	7
6.5—7.5	7	6	42	+2	4	24
7.5—8.5	8	3	24	+3	9	27
		$n=36$	$\Sigma m=186$			$\Sigma d^2_x=106$

[1] The assumed average x i.e. is a whole number approximating the actual average. That value which has the maximum frequency may also be taken as the assumed average, or the working mean as it is often called.

True Arithmetic Average or $a = \frac{\Sigma m}{n} = \frac{186}{36}$

$= 5{\cdot}17$ approximately

Let the Assumed Average or $x=5$; $(a-x)^2 = \left(\frac{186}{36} - 5\right)^2 = \frac{1}{36}$

Standard Deviation or $\sigma = \sqrt{\frac{\Sigma d^2 - n\ (a-x)^2}{n}}$

$= \sqrt{\frac{106 - 36\ (\frac{1}{36})}{36}} = 1{\cdot}71$ approx.

Standard Co-efficient of Dispersion $= \frac{\sigma}{a} = \frac{1{\cdot}71}{5{\cdot}17} = {\cdot}33$ approx.

Characteristics and Uses of Standard Deviation and its Co-efficient.

Standard deviation and its co-efficient possess all those properties which a good measure of dispersion should possess. The process of squaring the deviations eliminates negative signs, and thus makes mathematical manipulation of figures easy. Largely for this property, standard deviation has been used by biologists. It has not found its favourites among economists for two main reasons. Firstly, the squaring of deviations gives greater weight to extreme items than it does to those differing only slightly from the arithmetic average. This factor has hardly any value in economic studies, since the commercial or economic statistician is interested in the results of the modal class. Secondly, the computation of standard deviation requires considerably greater time and effort than that of mean deviation. For the businessman rapidity in the preparation of results is an important factor. Therefore, in economic and commercial studies there is a tendency to use mean deviation unless there is a particular reason for using the standard deviation.

But, standard deviation enjoys at least two decided advantages over the mean deviation. Firstly, it is, in general, less

affected by fluctuations of sampling, and secondly, it is more easily amendable to algebraical processes. These two properties make standard deviation useful for advanced work. Its use is, therefore, increasing for measuring variability. The standard deviation, for instance, has a special use in the computation of Karl Pearson's Co-efficient of Correlation, which will be discussed in a later Chapter.

Modulus.

It is another measure of dispersion based on the second moment of dispersion. It is generally represented by c. The formula used is : $c=\sqrt{\frac{2\Sigma d^2}{n}}$

Variance.

The quantity σ^2 is known as Variance.

Co-efficient of Variation.

According to Karl Pearson, who first suggested its use, co-efficient of variation is the 'percentage variation in the mean, the standard deviation being treated as the total variation in the mean.' In other words,

$$\text{Co-efficient of Variation or } v=100\,\frac{\text{Standard Deviation}}{\text{Arithmetic Average}}$$

$$=100\times\text{Co-efficient of Standard Deviation.}$$

Thus, in example 8, co-efficient of variation$=100\times{}^{\cdot}33=33$. Co-efficient of variation is a relative measure of dispersion and has come into use largely.

Quartile Deviation and its Co-efficient.

The measures of dispersion discussed under the First and the Second Moments of Dispersion take into account the deviations of all the items. **Quartile deviation,** also called the **Semi-Interquartile Range,** is based on the quartiles, and affords a general idea of the dispersion of a group without considering each particular item. The first and the third quartiles include between them the middle-half of the items of a group. If the

dispersion of this half could fairly represent the dispersion of the whole group, a simple method of measuring it will be :

$$\text{Quartile Deviation or } Q.\ D. = \frac{Q_3 - Q_1}{2},$$

where, Q_3 stands for 3rd or upper quartile, and Q_1 stands for 1st or lower quartile.

Quartile deviation is an absolute measure of dispersion. Its relative measure, that is, **Quartile Co-efficient of Dispersion,** is quartile deviation divided by the average of the two quartiles. Symbolically,

$$\text{Quartile Co-efficient of Dispersion} = \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Calculation of Quartile Deviation and its Co-efficient.

Example 9. In example 1, table 20, Q_3=Rs. 43·75 and Q_1=Rs. 40·25.

$$\therefore \text{ Quartile Deviation} = \text{Rs.} \frac{43{\cdot}75 - 40{\cdot}25}{2} = \text{Rs. } 1{\cdot}75$$

$$\text{Quartile Co-efficient of Dispersion} = \frac{43{\cdot}75 + 40{\cdot}25}{43{\cdot}75 - 40{\cdot}25}$$

$$= {\cdot}042 \text{ approximately.}$$

Upon the assumption that median lies half-way between the upper and the lower quartiles, we may observe that in our example.

$$\text{Median} = \frac{Q_3 + Q_1}{2} = \text{Rs. } \frac{43{\cdot}75 + 40{\cdot}25}{2} = \text{Rs. } 42$$

and the difference occuring on either side of it is Rs. 1·75. In other words, median is Rs. 42 and half the items are within the range Rs. 42±1·75.

Characteristics and Uses of Quartile Deviation and its Co-efficient.

Quartile deviation and its co-efficient are simple to comprehend and easy to compute. They are quite satisfactory if one

is concerned with the main body—the middle half—of the series and cares nothing about extreme variations. Quartile deviation has a serious drawback in that its value will be the same for those series whose quartiles are the same, whatever the distribution of the observations between the quartiles and beyond them. This is because it considers only the quartiles and not each particular item of the array. It is, therefore, not so sensitive as the mean and the standard deviations.

Choice of Measures of Dispersion.

A good measure of dispersion should possess some such qualities as an ideal average does. That is, it should be based on all the observations made, easily calculated, readily understood, affected very little by fluctuations of sampling, and amenable to algebraical treatment. The range, it has already been seen, is not a satisfactory measure, and its co-efficient, therefore, is not very much favoured by statisticians. Quartile deviation enjoys two advantages over the standard and mean deviations : It is easier to calculate and clearer in meaning. But, since it has no simple algebraical properties and is liable to be erratic, it is not good for any but the most elementary statistical work, where only a rough estimate is desired. It is, however, not unsatisfactory when the distribution of values in a series is fairly symmetrical. If the distribution lacks symmetry and there are great differences in frequency between successive values of items in the series, it is better to select measures which give each value its due weight. Such measures are the mean deviation and the standard deviation. Mean deviation is less troublesome to calculate than the standard deviation, but cannot be used for further mathematical operations. If in a given problem median suits the best, mean deviation would be a good measure of dispersion. But as arithmetic mean is the most commonly used average, standard deviation, which is invariably computed from the arithmetic average, is the most important and the best measure of dispersion. More so, because it is the least erratic, and is suitable for further algebraic manipulation. Its use is, therefore, recommended for cases where positive and comparatively precise results are desired.

Absolute and Relative Measures of Dispersion.

Two or more groups may be computed by stating their respective means and absolute measures of dispersion provided the means of the groups do not vary greatly in size and the groups are measured in the same unit. If the difference between the means is great, it is safer to compare their relative measures of dispersion, *i.e.*, co-efficients, rather than absolute measures. For example, if the production of a commodity in factory A for three successive years be 1250, 2000 and 2750 units respectively, and that of factory B for the same period runs in the order of 3250, 4000 and 4750 units, the mean deviation, 500 units, is exactly the same in both cases. Similarly are the ranges and the standard deviations identical. If only these absolute measures are compared, a fallacious conclusion that the variablity of production in the two factories is the same might be drawn. But, when the mean co-efficient of dispersion for the production of factory A, ·25, is compared with the mean co-efficient of dispersion for factory B, ·125, the difference in the degree of variability in the production of the two factories is made clear. According to the general rule that **the lower the co-efficient of dispersion** (mean, standard or quartile), **the smaller is the variability of the series,** or, in other words, the greater is its consistency, production in factory A is more variable than that in factory B. Co-efficients of dispersion, therefore, correct the wrong impression created by the absolute measures.

The absolute measures are concrete quantities, and should be stated in terms of the units of the variable (Rs. miles, inches, years, etc.). It is impossible to compare dispersions in two universes measured in different units—say, one measured in Rs. and the other in yards—through absolute measures of dispersion. For this reason also, as already pointed out co-efficients of dispersion are computed. They enable comparison between universes of different characters, since they are pure numbers.

Relation between Measures of Dispersion.

Mean, standard and quartile deviations all measure the same property, *viz.*, dispersion ; but they do it in different

ways. There does not exist a perfectly definite relation between them. Yet, for moderately symmetrical distributions the following relations are *approximately* true :

(1) Quartile Deviation $=0{\cdot}6745\sigma=\frac{2}{3}$ (Standard deviation), roughly.

(2) Mean Deviation $=0{\cdot}7979\sigma=\frac{4}{5}$ (Standard Deviation), approximately, when standard deviation is measured from the arithmetic average, a.

(3) $a+2\sigma$ or $a\pm3\sigma$ shall cover a majority of the observations of the group. (In example 5, a is 8 and σ is 3·066. Therefore, a majority of the items of the series would lie between 1·868 and 14·132 if the relationship given by $a\pm2\sigma$ were to hold good. Actually, the size of items in the given series vary between 2 and 14).

The above relations are not likely to hold good in cases where the number of observations is comparatively small. The formula (1) and (2) given above may, however, be used to determine one measure of deviation when another is known, provided the series is moderately symmetrical.

Lorenz Curve.

The graphic method can also be used for showing the dispersion of a group. The method adopted by Dr. Lorenz for the study of the distribution of wealth (the curve showing the distribution being designated after him as Lorenz Curve) is the best for the purpose. It is, in effect, a cumulative percentage curve, combining the percentage of items under review with the percentage of the factor (say, wealth) distribution among the items. In other words, it is an ogive curve[2] formed by cumulating the values of the items on each axis and reducing the values thus obtained to percentages of the total. Figure 1 on the next page illustrates the construction of the curve. If wealth were equally distributed among the people, the curve

[2] See the Chapter on Graphical Presentation.

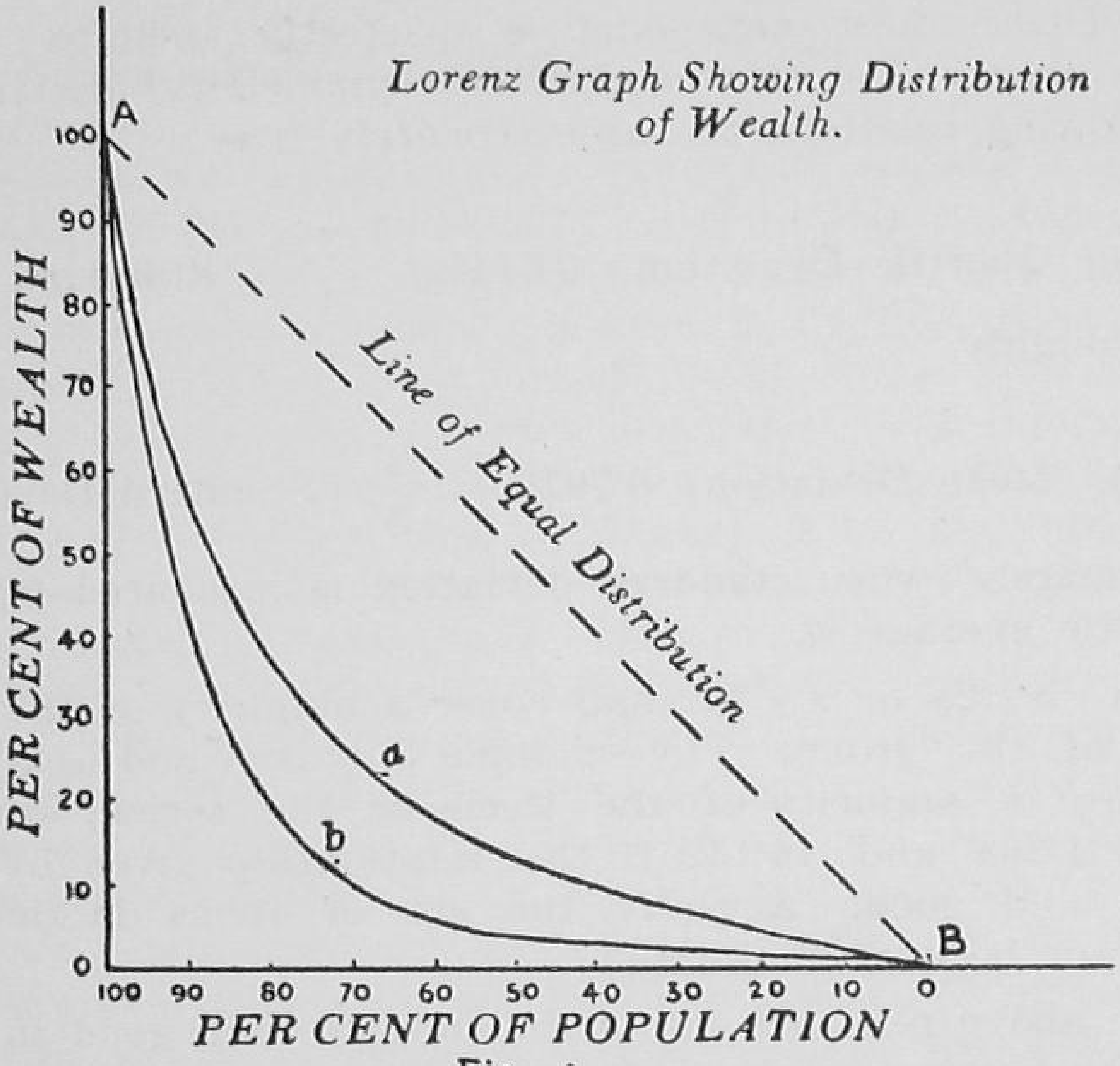


Fig. 1.

would be a straight line AB, connecting the two extremes of the scales. In practice, however, curves like *a* or *b* result. The less the distance between the curve showing actual distribution and the line of equal distribution, the greater is the homogeneity in the distribution of wealth or less is the dispersion. While, the further the curve of actual distribution from the line of equal distribution, the larger is the percentage of poor people and the greater is the concentration of wealth in the hands of a few millionaires. It is evident that dispersion in *b* is greater than that in *a* in Figure 1.

The method of calculating cumulative percentages is illustrated below where the given items are: (*a*) the value of turnovers in a certain trade and (b) the number of business-houses in that trade which had a given turnover.

Table 25A. *Turnovers & Number of Business-houses.*

Turnover exceeding (thousand rupees)	Number of Business-Houses	Cumulative Total of col. (*a*)	Col. (*c*) as Percentage of Total	Cumulative Total of col. (*b*)	Col. (*e*) as Percentage of Total
(*a*)	(*b*)	(*c*)	(*d*)	(*e*)	(*f*)
2	13	2	1.0	13	13.0
4	16	6	3.0	29	29.0
8	21	14	7.0	50	50.0
16	23	30	15.0	73	73.0
20	9	50	25.0	82	82.0
40	8	90	45.0	90	90.0
50	7	140	70.0	97	97.0
60	3	200	100.0	100	100.0

To draw the Lorenz Curve of turnover, Col. (*f*) (that is, cumulative percentages of business-houses) may be taken on the horizontal base and corresponding values of Col. (*d*) may be plotted on vertical scale. A curve like *a* or *b* in Figure 1 will then result. It will show the concentration of turnovers diverging from the average, for if the turnover of all business-houses were identical a straight line like AB in Figure 1 would show the equal distribution. (Note that in Figure 1 the arrangement of scale on the base line is reverse that on the vertical line).

The Lorenz Curve does not yield a numerical measurement of dispersion and is, in this respect, inferior to measures of dispersion. But its merit is that it affords a picture of dispersion at a glance. It should be used along with a co-efficient of dispersion when a detailed study of dispersion is desired. Lorenz Curve is very useful in such studies as the distribution of land, wages and income among the population of a country or the distribution of profits over different groups of businesses.

Practical Utility of Measures of Dispersion.

Measures of dispersion, it has already been indicated, supplement the information given by the mean. They are computed to determine the 'typical' character of an average. They

act as a useful check on drawing wrong conclusions from the comparison of statistical averages.

Measures of dispersion may also be computed for estimating the value of a series itself. For instance, the dispersion of time series affords a measure of the consistency or variability of the phenomenon to which the series relates. Determination of the degree of variability is, at times, very important in politico-economic problems. Violent fluctuations in production or trade naturally give a shock to the economic organism, which affects many people. Measures of dispersion enlighten us on the degree of these variations. Measures of dispersion are invaluable in the study of such problems as inequalities of income in a country, distribution of land among different units of agricultural community, wage fluctuations, etc. They enable comparison between different groups of phenomena, which is an important function that the Science of Statistics has to perform.

SKEWNESS

Skewness denotes the opposite of symmetry. As applied to frequency distribution, it indicates that the distribution of items in it is not symmetrical. Skewness relates to the shape of the curve of a frequency distribution, rather than to its size.

Tests of Skewness.

The presence of skewness or asymmetry in a given series can be tested in several ways. In a symmetrical series the mode, the median and the arithmetic average are identical. Therefore, skewness or lack of symmetry in a series is shown when these three averages do not coincide. It is further shown when the sum of the positive deviations from the median in a series is not equal to the sum of the negative deviations. It is also found to be existing when the quartiles, or pairs of deciles, are not equidistant from the median. It is also shown when at points of equal deviations on either side of the mode the frequencies are unequal. Lastly, if skewness is present in a frequency distribution, its graph will not give the normal, bell-shaped, symmetrical form ; rather, the base would be stretched to a greater extent on one side than on the other. In curves which are not far away from being symmetrical the median

usually covers over two-thirds of the distance travelled by the arithmetic average from the mode (See Figure 2). Therefore, approximately,

$$M = Z + \frac{2}{3}(a - Z)$$

$$\text{or, } Z - M = \frac{2}{3}(Z - a)$$

where, M stands for median, Z for mode and a for mean. In a skew curve mode, median and arithmetic average normally occur in sequence, the last being pulled away the largest in the direction in which the curve is skewed.

We shall apply the above tests to the following example.

Table 26. *Calculation of Arithmetic Average, Median, and Mean and Standard Deviations.*

1	2	3	4	5	6	7	8	9
Size of item m	Frequency f	Cumulative frequency cf	Total size of item mf	Dev. from the median and mode d_m d_z	Total Dev. from median and mode fd_m fd_z	Deviation from mean d	Square of deviation from mean d^2	Total square of deviation from mean fd^2
3	1	1	3	−4	4	−4.1	16.81	16.81
4	2	3	8	−3	6	−3.1	9.61	19.22
5	3	6	15	−2	6	−2.1	4.41	32.23
6	5	11	30	−1	5	−1.1	1.21	6.05
7	8	19	56	0	0	−0.1	.01	.08
8	6	25	48	+1	6	0.9	.81	4.86
9	4	29	36	+2	8	1.9	3.61	14.44
10	2	31	20	+3	6	2.9	8.41	16.82
11	1	32	11	+4	4	3.9	15.21	15.21
		$n=32$	$\Sigma m=227$	−10 +10	45	−10.5 + 9.6		Σd^2 =106.7

First Test : The arithmetic average is $\frac{227}{32}$ or 7·1, the median, that is, value of 16·5th item is 7, and the mode is also 7. The values of mean, median and mode, therefore, do not exactly coincide. The curve is not perfectly symmetrical.

Second Test : Deviations from the median in column 5 show fine symmetry. The negative sum of deviations,—10, is equal to the positive sum, +10, leading us to think that the curve is symmetrical.

Third Test : $Q_3 - M = 8 - 7 = 1$; $M - Q_1 = 7 - 6 = 1$. The two quartiles happen to be equi-distant from the median. Therefore, the curve of the series appears to have a symmetrical form.

Fourth Test : We take equal deviations on either side of the mode and compare the frequencies centering round them. 6 and 8 are equidistant from the mode which is 7. But the frequencies against 6 are 5 and against 8 are 6. Therefore, the series is not perfectly symmetrical.

The first and the fourth tests indicate that the shape of the curve of the given frequency distribution lacks symmetry ; while, the second and the third tests show that the curve might be symmetrical. The second and the third tests do not always provide the correct answer. We may, therefore, conclude that the curve of the given series falls slightly short of perfect symmetry. *'By how much does the curve lack symmetry'* ? may be the natural question. For answering it, we must reduce skewness to numerical quantity. This brings us to the measures of skewness.

Measures of Skewness.

The first measure of skewness, and the simplest too, is based on the fact that in a skew distribution the mean and the mode are drawn widely apart. The larger the distance that the mean (a) is pulled beyond the mode (Z), the greater is the degree of skewness.

The second measure of skewness is based on the fact that in a skew curve the median does not lie half-way between the quartiles, the quartile nearest to the stretched base being pulled in that direction more than the other quartile.

But these measures of skewness should be reduced to **co-efficients of skewness** for just the same reasons as measures of dispersion are reduced to co-efficients. In computing the co-efficients of dispersion the measures of absolute dispersion are divided by the average used. Average would not be the proper divisor here, for the question now is not how much the curve is asymmetrical in proportion to values of the items of the series, but how much more the items on one side deviate than they do on the other. Therefore, measures of skewness shall be divided by the related measures of dispersion. On these principles are based the measures and co-efficients of skewness discussed below.

First Measure and Co-efficient of Skewness.

Measure of Skewness$=a-Z$, *i.e.*, the difference between mean and mode,

$$\text{Co-efficient of skewness or } j=\frac{a-Z}{\delta_z} \ldots\ldots\ldots\ldots (A)$$

$$\text{or } j=\frac{a-Z}{\delta} \ldots\ldots\ldots\ldots (B)$$

And if the mode is ill-defined, the numerator may be substituted by the difference between mean and median,, so that,

$$j=\frac{a-M}{\delta_m} \ldots\ldots\ldots\ldots (C)$$

Karl Pearson has given a formula in which standard deviation is employed as the denominator rather than mean deviation used above, so that,

$$j=\frac{a-Z}{\sigma} \ldots\ldots\ldots\ldots (D)$$

And if the mode is ill-defined, then on the basis of the relationship between median, mode and mean pointed out above[3], formula (D) may be modified as below :

$$j=\frac{3\,(a-M)}{\sigma} \ldots\ldots\ldots (E)$$

3 See page 185.

First Test : The arithmetic average is $\frac{227}{32}$ or 7·1, the median, that is, value of 16·5th item is 7, and the mode is also 7. The values of mean, median and mode, therefore, do not exactly coincide. The curve is not perfectly symmetrical.

Second Test : Deviations from the median in column 5 show fine symmetry. The negative sum of deviations,—10, is equal to the positive sum, +10, leading us to think that the curve is symmetrical.

Third Test : $Q_3 - M = 8 - 7 = 1$; $M - Q_1 = 7 - 6 = 1$. The two quartiles happen to be equi-distant from the median. Therefore, the curve of the series appears to have a symmetrical form.

Fourth Test : We take equal deviations on either side of the mode and compare the frequencies centering round them. 6 and 8 are equidistant from the mode which is 7. But the frequencies against 6 are 5 and against 8 are 6. Therefore, the series is not perfectly symmetrical.

The first and the fourth tests indicate that the shape of the curve of the given frequency distribution lacks symmetry ; while, the second and the third tests show that the curve might be symmetrical. The second and the third tests do not always provide the correct answer. We may, therefore, conclude that the curve of the given series falls slightly short of perfect symmetry. *'By how much does the curve lack symmetry' ?* may be the natural question. For answering it, we must reduce skewness to numerical quantity. This brings us to the measures of skewness.

Measures of Skewness.

The first measure of skewness, and the simplest too, is based on the fact that in a skew distribution the mean and the mode are drawn widely apart. The larger the distance that the mean (a) is pulled beyond the mode (Z), the greater is the degree of skewness.

The second measure of skewness is based on the fact that in a skew curve the median does not lie half-way between the quartiles, the quartile nearest to the stretched base being pulled in that direction more than the other quartile.

But these measures of skewness should be reduced to **co-efficients of skewness** for just the same reasons as measures of dispersion are reduced to co-efficients. In computing the co-efficients of dispersion the measures of absolute dispersion are divided by the average used. Average would not be the proper divisor here, for the question now is not how much the curve is asymmetrical in proportion to values of the items of the series, but how much more the items on one side deviate than they do on the other. Therefore, measures of skewness shall be divided by the related measures of dispersion. On these principles are based the measures and co-efficients of skewness discussed below.

First Measure and Co-efficient of Skewness.

Measure of Skewness $=a-Z$, *i.e.*, the difference between mean and mode,

$$\text{Co-efficient of skewness or } j=\frac{a-Z}{\delta_z} \text{ ....(A)}$$

$$\text{or } j=\frac{a-Z}{\delta} \text{ ....(B)}$$

And if the mode is ill-defined, the numerator may be substituted by the difference between mean and median,, so that,

$$j=\frac{a-M}{\delta_m} \text{ ....(C)}$$

Karl Pearson has given a formula in which standard deviation is employed as the denominator rather than mean deviation used above, so that,

$$j=\frac{a-Z}{\sigma} \text{ ....(D)}$$

And if the mode is ill-defined, then on the basis of the relationship between median, mode and mean pointed out above[3], formula (D) may be modified as below :

$$j=\frac{3\,(a-M)}{\sigma} \text{ ....(E)}$$

[3] See page 185.

The above co-efficients yield a pure number, and are, therefore, independent of the units in which the variable is measured, and, secondly, they shall give a zero for symmetrical series. These are the two properties which a good measure of skewness should possess. It is why the above methods of measuring skewness are termed ideal or standard methods.

According to these formulæ the following are the results of our series. The measure of skewness $=a-Z=7{\cdot}1-7=+{\cdot}1$

$$j=\frac{a-Z}{\delta_z}=\frac{\cdot 1}{1{\cdot}406}=+{\cdot}071 \qquad j=\frac{a-Z}{\delta}=\frac{\cdot 1}{1{\cdot}425}=+{\cdot}0702$$

$$j=\frac{a-M}{\delta_m}=\frac{7{\cdot}1-7}{1{\cdot}406}=+{\cdot}071 \qquad j=\frac{a-Z}{\sigma}=\frac{\cdot 1}{1{\cdot}825}=+{\cdot}0548$$

$$j=\frac{3(a-M)}{\sigma}=+{\cdot}1644$$

In theory there is no limit to the values of co-efficients yielded by the formulæ A, B, C, & D, and this is a slight drawback. In practice the results are rarely very high, and for moderately asymmetrical curves they are usually less than unity. The co-efficient yielded by the formula E lies between the limits -3 and $+3$. In practice it rarely approaches that limit.

Since none of the above formulæ yields zero co-efficient of skewness for our series, we are led to conclude that the curve is skew. But, since the value of the co-efficient is very small, we may add that the degree of skewness is very slight. Skewness, it should be noted, is positive in this case, *i.e.*, the mean is greater than the mode.

Second Measure and Co-efficient of Skewness.

The Measure of Skewness $=(Q_3-M)-(M-Q_1)$

$$=Q_3+Q_1-2M,$$

where, Q_3 stands for upper quartile, Q_1 for lower quartile and M for median.

The Co-efficient of Skewness, or $j=\dfrac{(Q_3-M)-(M-Q_1)}{(Q_3-M)+(M-Q_1)}$

$$=\frac{Q_3+Q_1-2M}{Q_3-Q_1}\ldots\ldots\ldots(\text{F})$$

This co-efficient is also a pure number, and is zero for symmetrical distributions. Its result varies from -1 to $+1$. The fact whether the particular value of a co-efficient is significant or not is a matter of experience. It may, however, be suggested that ·1 denotes a moderate degree and ·3 a considerable degree of skewness.

According to the above formulæ, in our series, table 26, page 185,

$$\text{Measure of Skewness} = Q_3 + Q_1 - 2M$$
$$= 8 + 6 - 14$$
$$= 0$$

$$\text{Co-efficient of Skewness} = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$
$$= \frac{0}{2}$$
$$= 0$$

The second measure of skewness and its co-efficient suggest that the curve of the frequency distribution in table 26 is symmetrical—a suggestion not warranted by the first measure and co-efficient of skewness. This is the weakness of the second measure and co-efficient. It is the same weakness as that possessed by quartile co-efficient of dispersion——that is, it fails to take into account the size of the extreme variations, since it is concerned only with the quartiles and the median. This limitation should be clearly borne in mind before results yielded by this formula are relied upon. This co-efficient is simple and easy to calculate and is sufficiently reliable in those studies in which extreme instances are considered unimportant. It is a rather rough-and-ready measure and might be used where quartile deviation is being used as a measure of dispersion. Where comparatively greater accuracy is required, Karl Pearson's Co-efficient of Skewness should be employed.

Positive and Negative Skewness.

Skewness can be positive as well as negative. If the arithmetic average is greater than the mode or the median, skewness is positive. If it is less, skewness is negative. When

skewness is positive, mean would travel to the right of the mode in the curve. It would be to the left of the mode, when skewness is negative. In our example, the answer is positive. Positive and negative skewness are symbolised by plus and minus signs respectively.

Figure 2, below, illustrates slight *positive* skewness. It also shows the positions of the mode, the median and the arithmetic mean, indicated by *Z, M,* and *a* respectively, in an ideal moderately asymmetrical distribution: the median travels about 2/3rds the distance travelled by the mean from the mode. It is also clear that the mode, the median and the mean occur in sequence in a skew curve, and the mean is pulled away the largest in the direction in which the curve is skewed. It may

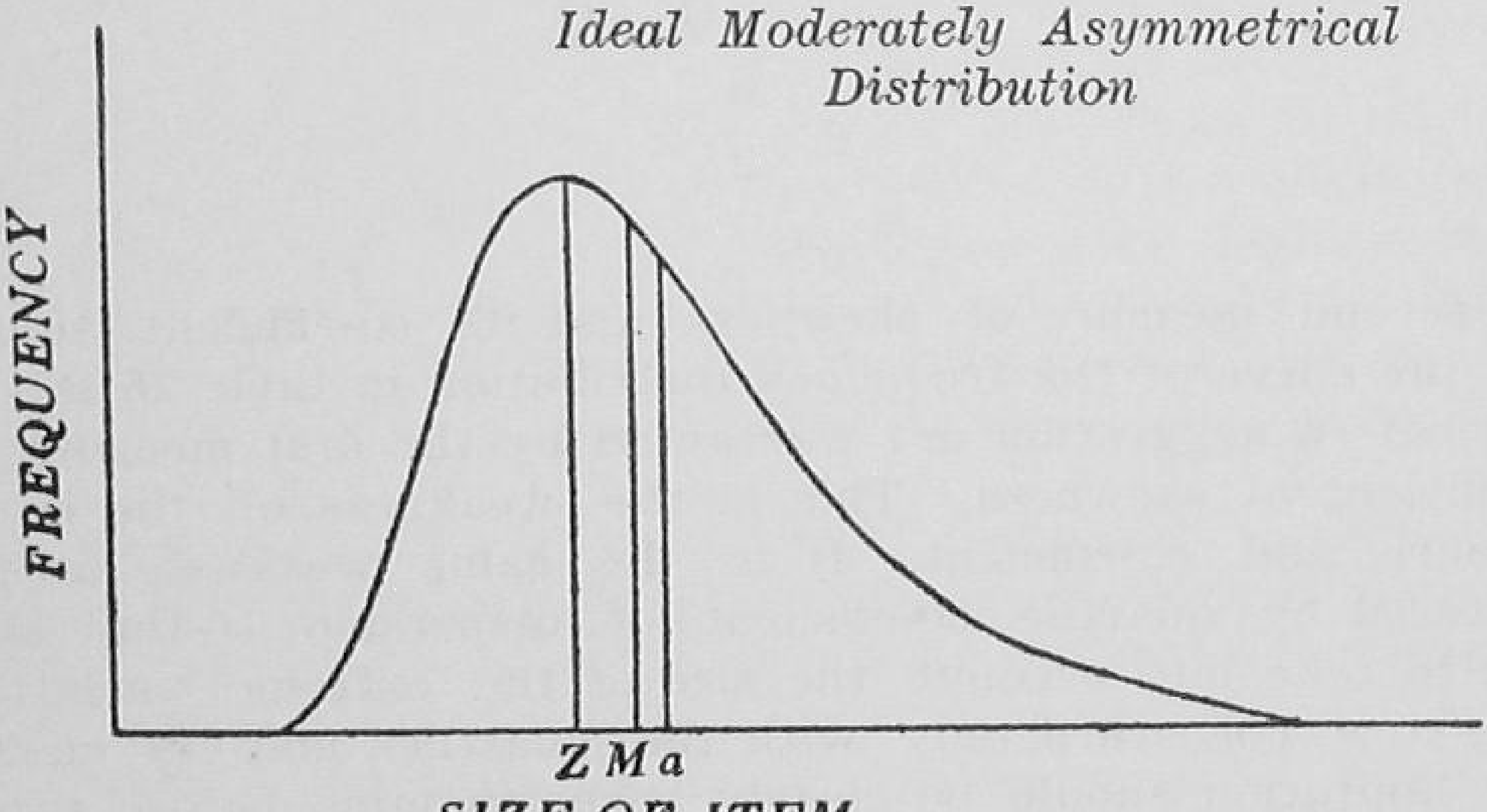


Fig.2.

also be noted that the median bisects the area of the curve (called histogram). Mode will remain unaffected by the addition of a few more items, but the median and the mean will be deflected.

Dispersion and Skewness Contrasted.

Measures and co-efficients of dispersion, respectively, indicate the absolute and relative differences between the individual items of the series and an average taken as the standard. They do not, however, show the extent to which deviations cluster above or below the average selected.

Measures of skewness, on the other hand, show the extent to which distributions are pulled away, or distorted, from the ideal, symmetrical curve. In a symmetrical curve the mode, the median and the mean coincide ; in an unsymmetrical curve they do not. Measures of skewness have two functions to perform : firstly, they indicate the direction of asymmetry through their positive and negative character ; secondly, they measure the amount of asymmetry in absolute or relative terms through the value obtained for the measure or the co-efficient.

The theory of skewness is more important in biological studies and other studies depending more or less upon the laboratory experiments than in economic and social investigations. In social and economic inquiries a perfectly symmetrical distribution is an exception, and a large degree of skewness is generally expected.

It is interesting to note the important part that median and quartiles play in statistics. Three characteristics of a group can be studied through them alone, the median locating the central position, quartile deviation showing dispersion, and the second measure of skewness showing skewness.

EXERCISES

(1) Discuss the various ways in which the differences in the characteristics of frequency distribution are generally measured.

(B. Com., Luck, 1937).

(2) What is meant by Dispersion ? What are the methods of computing measures of dispersion ? Illustrate the practical utility of such measures.

(M. Com., Alld., 1945).

(3) Explain the meaning of the term 'Dispersion', and distinguish between absolute and relative measures of dispersion.

(B. Com., Alld., 1946).

(4) Explain the various methods of describing the scatter of a frequency distribution and say what you know as to the relative worth of the resulting measures.

(B. Com., Nag., 1944).

(5) Define carefully mean deviation, standard deviation and quartile deviation of any given distribution. In what problems should each be used ? In a non-symmetrical distribution how would you find the skewness ?

(M.A., Alld., 1940).

(6) What is meant by Skewness ? How does it differ from dispersion ? What is the practical utility of these measures ?

(M.A., Pb., Apr., 1952).

(7) Frequency distributions may either differ in the numerical size of their averages though not necessarily in their formations, or they may have the same values of their averages yet differ in their respective formations.

Explain and illustrate how the measures of dispersion afford a supplement to the information about frequency distributions furnished by the averages.

(M. Com., Raj., 1952).

(8) Calculate the Co-efficient of Variation of the data given in exercise 15, Chapter IX.

(B. Com., Alld., 1947).

(9) Calculate Karl Pearson's co-efficient of skewness from the following data :

Marks	No. of Students
Above 0	150
,, 10	140
,, 20	100
,, 30	80
,, 40	80
,, 50	70
,, 60	30
,, 70	14
,, 80	0

(B. Com., Alld., 1953).

(10) (*a*) Describe briefly *two* of the measures used for describing the dispersion of a statistical series about its mean. The frequency distribution of the earnings of motor bus drivers, general medical practitioners and stock exchange speculators in rupees per annum have standard deviations A, B and C. Arrange A, B and C in ascending order of magnitude.

(*b*) Describe the difference in the earnings of balers and boiler-cleaners from the following:—

	Balers	Boiler Cleaners
	Rs. a. p.	Rs. a. p.
Daily mean earnings ..	1 2 1	1 2 1
Lower Quartile ..	0 15 0	0 14 0
Median	1 2 0	0 15 8
Upper Quartile ..	1 4 9	1 2 0

(B. Com. Hons., Andhra, 1942).

(11) Calculate the coefficient of variation of the following monthly incomes of twenty families given below in rupees:—

2,000; 35; 400; 15; 40; 1,500; 300; 6; 90; 250; 20; 12; 450; 10; 150; 8; 25; 30; 1,200; 60.

(B. Com., Alld., 1941).

(12) Find the Arithmetic Average, the First Moment of Dispersion, and the Standard Deviation from the data in the following series:—

Size of item	Frequency
3— 4 ..	3
4— 5 ..	7
5— 6 ..	22
6— 7 ..	60
7— 8 ..	85
8— 9 ..	32
9—10 ..	8

(B. Com., Alld., 1942).

(13) The following table shows the number of workers in two factories whose weekly earnings are given in column (1). Determine the mean values of weekly earnings and standard deviation in both factories.

Range of weekly earnings	Number of workers in	
	Factory A	Factory B
4—6	74	71
6—8	376	379
8—10	304	303
10—12	110	112
12—14	18	18
14—16	0	1
16—18	9	3
18—20	9	9
20—22	0	4
TOTAL	900	900

(M.A., Cal., 1936).

(14) Discuss the usefulness of the coefficient of variation. Calculate the coefficient of variation of height distribution from the following table:—

Height in inches	Number of men	Height in inches	Number of men
57	2	68	1,230
58	4	69	1,063
59	14	70	646
60	41	71	392
61	83	72	202
62	169	73	79
63	394	74	32
64	669	75	16
65	990	76	5
66	1,223	77	2
67	1,329	..	..

(M.A., Alld., 1951).

(15) The following table shows the size of certain units sampled out in various coal-fields in India. In which area is the average size of the coal-mining unit the largest ?

No. of Persons employed	Number of units in Jharia	Raniganj	Madhya Pradesh
Below 50	3	3	1
50— 100	12	4	1
100— 200	18	10	8
200— 300	16	12	3
300— 400	14	8	5
400— 500	6	9	1
500—1000	19	13	6
1000 & above	13	9	7

(M.A., Pb., Apr., 1952).

(16) Calculate the mean, the median and the standard deviation of the following distribution :

Marks obtained	10-	15-	20-	25-	30-	35-	40-	45-	50-	55-	60-
No. of Candidate	10	23	29	35	40	52	70	60	12	9	3

(M.A., Patna, 1944).

(17) Below are given the profits earned by two groups of business, A and B. Use a graphic method to compare both the groups, with a view to studying how the profits earned in each group deviated from the average and from one another.

(M. Com., Alld., 1947).

Total amount of profits earned by Companies in each division	No. of Companies in each division	
	Group A	Group B
Rs.		
1,500	15	3
6,250	27	47
15,000	33	65
21,000	35	35
26,250	37	35
37,500	43	32
42,500	25	15
1,00,000	35	18

(M. Com., Alld., 1947).

(18) Calculate the mean deviation from the following data. What light does it throw on the social conditions of the community?

Difference in age between husband and wife in a particular community

Difference in years	Frequency
0— 5	449
5—10	705
10—15	507
15—20	281
20—25	109
25—30	52
30—35	16
35—40	4

(B. Com., Bom., 1936).

(19) What do you understand by the term 'dispersion'? Name and describe two standard measures of dispersion.

The index numbers of prices of cotton and coal shares in 1942 were as under:—

Month	Index number of prices of cotton shares	Index number of prices of coal shares
January	188	131
February	178	130
March	173	130
April	164	129
May	172	129
June	183	129
July	184	127
August	185	127
September	211	130
October	217	137
November	232	140
December	240	142

Which of the two shares do you consider more variable in price?

(M.A., Agra, 1944).
(B. Com., Alld., 1952).

(20) Calculate the mean and the standard deviations of the following figures and state the percentage of cases which lie outside the mean at distances $a \pm \sigma$, $a \pm 2\sigma$, $a \pm 3\sigma$, where σ stands for the standard deviation.

115, 117, 121, 125, 116, 120, 118, 117, 119, 116,
122, 124, 123, 118, 120, 118, 126, 127, 122, 123,

(21) *Summary of Receipts and Passengers of a certain Motor Bus Co.*

Year		Receipts	Passengers
1925	..	2,354	50,010
1926	..	2,780	61,060
1927	..	3,011	70,000
1928	..	3,020	70,110
1929	..	3,541	83,001
1930	..	4,150	91,100
1931	..	5,000	100,000

From the foregoing data, find out one measure of dispersion and state whether the variation in receipts is greater than that in passengers.

(B. Com., Alld., 1932).

(22) The following 22 price relatives are available for the construction of index number of prices:

53, 61, 61, 64, 64, 70, 71, 73, 76, 78, 81, 85, 85,
93, 94, 96, 97, 101, 102, 139, 143 and 144.

Regarding these as a statistical group, calculate their mean, median and co-efficient of dispersion.

Will you select the mean or the median as the appropriate average for the index number in question? Give reasons for your selection.

(M.A., Pb., Sept. 1952).

(23) Find the mean, median, and modal ages at first child births of married women from the table given below, and calculate the standard deviation and the co-efficient of variation:—

Age at the birth of first child	No. of married women
13	37
14	162
15	343
16	390
17	256
18	433
19	161
20	355
21	65
22	85
23	49
24	46
25	40

(M.A., Patna, 1943).

(24) Calculate the Standard Deviation of the following data with regard to 2,298 families in the U. K.

Number of persons in the family	1	2	3	4	5	6	7	8	9	10	11	12	Total
Number of families	166	552	580	433	268	148	77	41	20	8	5	1	2298

(M.A., Alld., 1942).

(25) The following are the rents of 18 houses in a certain locality:—

Rs.	a.	Rs.	a.
6	8	6	4
5	0	3	0
5	4	9	0
5	8	4	8
5	4	4	0
4	12	5	0
4	0	3	12
5	0	5	0
4	8	3	0

Calculate the mean deviation of this group.

(B. Com., Luck., 1930).

(26) Draw Lorenz Curve from the following data to show the extent of dispersion graphically.

Amount of Profits Rs.	No. of Companies
600	6
2,500	11
6,000	13
8,400	14
10,500	15
15,000	17
17,000	10
40,000	14

(M.A., Pb., Sept. 1952).

(27) The following table gives the number of finished articles turned out per day by different numbers of workers in

a factory. Find the mean value and 'standard deviation' of the daily output of finished articles, and explain the significance of 'standard deviation'—

Number of articles	Number of workers	Number of articles	Number of workers
18	3	23	17
19	7	24	13
20	11	25	8
21	14	26	5
22	18	27	4

(B. Com., Cal., 1937).

(28) Write short notes on
(1) Dispersion.
(2) Standard deviation.

Calculate the standard deviation from the following data:

Size of item	Frequency
6	3
7	6
8	9
9	13
10	8
11	5
12	4
TOTAL	48

(B. Com., Bom., 1936).
(B. Com., Nagpur, 1944).

(29) Find out Karl Pearson's coefficient of skewness from the following figures:—

Income in Rupees	Number of Persons
1 and not exceeding 9	75
10 ,, ,, ,, 19	100
20 ,, ,, ,, 29	302
30 ,, ,, ,, 39	603
40 ,, ,, ,, 49	452
50 ,, ,, ,, 59	63
60 ,, ,, ,, 69	15

(M. Com., Alld., 1948).

(30) The following table gives the area under cultivation, production, net export, and price of wheat in the U. S. A.:—

Year	Area under cultivation (millions of Acres)	Production (millions of Bushel)	Net export (millions of Bushel)	Price per Bushel in Cents.
1890	36.7	449	109	84
1895	38.9	542	130	51
1900	49.2	599	221	62
1905	46.3	706	101	75
1910	45.8	625	70	91
1915	60.3	1008	240	96
1920	62.3	843	313	183
1925	52.4	669	93	144
1930	62.6	886	112	67
1935	51.3	626	..	83

Compare the coefficient of variation of area under cultivation and net export of wheat for the whole period.

(M.A., Patna, 1941).

(31) Find out standard deviation and a coefficient of skewness for the given distribution:—

Variable	Frequency
0— 5	2
5—10	5
10—15	7
15—20	13
20—25	21
25—30	16
30—35	8
35—40	3
	75

(B. Com., Alld., 1951).

(32) Calculate the standard deviation from the data given in exercise 12, Chapter IX.

(B. Com., Nag., 1945).

(33) In any two samples, where the variates x_1 and x_2 are measured in the same units

$$n_1=36 \; ; \; \Sigma x_1^2=49428$$

$$n_2=49 \; ; \; \Sigma x_2^2-71258$$

Compute the values of the standard deviations of the two samples. What additional information is required to calculate the coefficient of variation of the above two samples ? Indicate the uses of such a coefficient.

(B. Com., Luck., 1943).

(34) What relationship generally subsists between different measures of dispersion in a moderately symmetrical series ?

If in a series which is not highly skewed the mean deviation is 9·8 feet, what would be the approximate value of its standard deviation ?

(M.A., Pb., Sept., 1951).

(35) The standard deviation of scores for a given group of pupils on an arithmetic test is 28 and on a spelling test is 18. It is concluded that this group is relatively more variable in arithmetic than in spellings. Do you agree ?

(M.A., Pb., Apr., 1951).

(36) Write short notes on :

(*a*) Lorenz Curve,
(*b*) Skewness,
(*c*) Averages of Second Order,
(*d*) First Moment of Dispersion.

(37) Explain what you understand by the terms

(*i*) Coefficient of Quartile Deviation,
(*ii*) Coefficient of Mean Deviation, and
(*iii*) Coefficient of Standard Deviation.

Discuss their relative merits in analysing a given data.

(M.A., Pb., 1950).

(38) Explain the meaning and significance of 'skewness'. Which of the following two distributions is more skew ?—

Distribution of weekly index numbers of cost of living in Bombay in 1942		Distribution of weekly index numbers of cost of living in Bombay in 1943	
Index No.	No. of weeks	Index No.	No. of weeks
140—150	5	200—210	10
150—160	10	210—220	10
160—170	20	220—230	10
170—180	9	230—240	8
180—190	6	240—250	7
190—200	2	250—260	7

Represent the distributions graphically.

(M.A., Agra, 1945).

(39) Sketch a skew frequency distribution showing the approximate positions of the arithmetic mean, the median and the mode.

(B. Com., Nag., 1942).

(40) The marks obtained by students of Classes A and B are given below. Give as much information as you can regarding the composition of the classes in respect of intelligence.

Marks obtained	No. of students in class A	No. of students in class B
5—10	1	5
10—15	10	6
15—20	20	15
20—25	8	10
25—30	6	5
30—35	3	4
35—40	1	2
40—45	0	2
45—50	0	0

(B. Com., Alld., 1952).

(41) The following table gives the marks obtained by a batch of 15 candidates in an examination in History, Politics and Economics:—

Roll No.	Marks obtained in History	Politics	Economics
1	42	46	35
2	24	20	25
3	38	41	44
4	35	43	50
5	30	25	15
6	45	54	57
7	58	47	55
8	50	36	40
9	40	30	20
10	62	61	64
11	55	50	42
12	54	63	60
13	52	45	62
14	47	56	54
15	43	58	52

In which subject is the level of knowledge highest ? Give reasons.

(M. Com., Alld., 1950).

CHAPTER XI

INDEX NUMBERS

An Index Number is a number which indicates the level of a certain phenomenon at any given date in comparison with the level of the same phenomenon at some standard date. It offers a device for estimating the relative changes of a variable in cases where measurement of its actual changes is inconvenient or impossible. If we want to measure the changes from one period to another in a factor, for change may not be capable of direct measurement. But, an evidence of it may be had from a measurement of the quantities *influenced* by the factor under consideration. These measurements may be expressed in different units. Therefore, their movements will not be directly comparable. To make them comparable, we may reduce the changes to a common denominator. We may, therefore, express them as percentages of similar measurements for a selected date. When so expressed, the percentages shall form a group. Each one of these percentages shall throw some light on the incommensurable, hidden factor about which we desire information. If we take an average of all these percentages, it would afford an approximate idea of the change in the factor in question. This average is called Index Number. Thus, averages linked with percentages constitute the whole basis upon which is raised the superstructure of a simple device of comparing factors which are not *directly* comparable.

Let us take an example. Suppose we are concerned with measuring the general changes in the price level. These changes are not directly measurable. Evidence of them can be seen in changes in the prices of different commodities. Quotations of these prices shall be available in different units, *e.g.* of wheat in Rs. per maund, of cotton in Rs. per bale, of petrol in Rs. per gallon. They are not directly comparable. If we plot them on a graph paper, reliable conclusions with regard to their movements will hardly be drawn. To make them comparable, we may express them as percentages of corresponding prices of

some selected date. These percentages, relating to different commodities for the same date, shall constitute a group. Each of these percentages shall reflect, in one way or the other, the change that has taken place in the general price level. When we compute an average of these percentages, the resulting average would show an approximate general change in the level of prices from the standard date to the date under consideration. This average is called the Wholesale Price Index Number, or the General Index Number.

Index numbers are not used for measuring changes in the general level of prices alone. They are as well employed to measure movements of wages, employment, cost of living, sales, production, exports, imports, investment, business activity, shares and stocks and a multitude of other phenomena over a period of time. In fact, where an attempt is made to bring to light what is enshrouded in complex variations of the items of a time series, they are invaluable to use.

General Characteristics of Index Numbers.

Movement in prices is a matter of general economic interest. To a layman a rupee is just a rupee, but changes in its purchasing power are very often a nuisance. General index numbers provide a measure of change in purchasing power. When the price of a single commodity changes, the reason for it may be found in a change in the supply of that commodity without a corresponding change in its demand or in a change in its demand without a corresponding change in its supply. And, if the price of a commodity which is a substitute for another changes, the reason for it may, in addition, be attributed to a change in the supply or demand of the substitute. But, when the prices of two, or a number of, unrelated commodities, say brassware and cloth, rise or fall together, several reasons may be advanced to explain the situation, but the real one may lie in an alteration in the measuring rod itself. This measuring rod is money, and its value, unlike that of other measuring instruments such as the foot-scale, the maund and the yard-stick, changes with supply of and demand for it. Change in the value of money implies change in prices.

According to the Quantity Theory of Money, a fall in the value of money is the same thing as a rise in the general level of prices, or *vice versa*. But, if change in the value of money were the only cause of change in prices, the prices of all commodities would show nearly the same proportional rise or fall, and, therefore, it would be easy to say in what direction the value of money or its purchasing power moved by looking at the change in the price of any single commodity. But as we know, change in the value of money is only one of the several causes, and it may be so intermingled with other causes like fluctuations of demand for and supply of goods that it may be difficult to say to what extent the value of money has changed. This complex phenomenon is simplified through the device of Price Index Numbers.

As Bowley puts it, "......the method of index numbers is at once applicable to the disentanglement of that which is common to the whole group from those variations which are special to individual items."* Prices, for instance, form a group but prices of different commodities constitute a heterogeneous group. This heterogeneous group is governed by two types of forces: general and special. The general force acts on all quantities of a group in the same sense, that is, it tends to increase all or decrease all. The special forces act on one or more of the quantities ; some of them tend to increase and others to decrease the quantities they affect ; some of the special forces, therefore, tend to increase the average, others tend to diminish it. The general force produces biased change; it will have a cumulative effect entirely in one direction—either towards increasing or towards decreasing the average. Special forces produce unbiased changes; these changes tend to cancel one another; their effect on the average is negligible. The change in the average will, then, show the influence of the general forces only. Thus, index numbers disentangle the commonness of rise or fall (that is, the effect of the general forces) from the effect of the special forces affecting the given quantities of a group.

* *Elements of Statistics*, 1920.

CONSTRUCTION OF WHOLESALE PRICE INDEX NUMBERS.

We have seen the necessity of constructing Index Numbers of Prices. The technique of their construction, or, for the matter of that, of the construction of any index number, involves the consideration of the following major problems:—

1. The selection of items to be included; their number and quotations.
2. The choice of the base period and calculation of price relatives.
3. The type of average to be used.
4. The system of weighting to be adopted.

The choice of commodities, base period, average and weights will largely depend on the purpose which the index number to be constructed is to serve. If the object is to measure the behaviour of *sensitive* prices, only those commodities will find place in the index which react quickly to market changes. This was the case with the Wholesale Price Index Number of 23 articles started by the Economic Adviser to the Government of India during the World War II and given up at the end of 1947. A *general purpose* index, on the other hand, will include a far larger number of commodities representative of the different sectors of the country's economy. This is true of the Economic Adviser's current Index Number of Wholesale Prices started at the beginning of 1947. If the purpose of an index number series is to measure price changes in comparison with a particular year or period, the question of choosing the base year does not arise; the base year is already given. If an index is to measure the price trends of only a particular commodity, the need for weighting does not arise. Thus it is important to first ascertain the object of the index, before proceeding with or arranging for its construction.

SELECTION OF ITEMS

Nature of Commodities.

The general wholesale price index number is based on the wholesale prices of commodities exchanged in the market. But, it is neither possible to include nor to obtain regular price

quotations of *all* the commodities that are bought and sold in the various markets of a country. Therefore, a sample of commodities has to be taken so as to bring down their number to a manageable limit. The commodities sampled out should be:

(a) representative of the tastes, habits and customs of the people;

(b) easily recognisable and describable; and

(c) unlikely to vary in quality, or those whose varieties have been standardised or graded or assigned trade descriptions.

For a *sensitive* index, the commodities chosen should be specially sensitive to fluctuations in business conditions. For a *general-purpose* index, the sample should include representatives of prices entering the general price system.

Number of Commodities.

There is no hard and fast rule for deciding the number. It is a fell recognised principle of statistical theory that the larger the size of a sample, the greater is the tendency for errors to compensate one another. But there are also the questions of administrative convenience to handle the data, as also questions of expense and delay, which are bound to assume importance as the number of commodities increases. As a general working rule, therefore, a reasonable number of commodities should be chosen, consistent with economy, simplicity and accuracy of measurement of price behaviour.

Sensitive indices cover smaller number of commodities than *general-purpose* index numbers. Economic Adviser's sensitive index—already referred to—based on 23 commodities in India, the index of 15 commodities compiled by the Bank of England and that of 20 basic commodities constructed by the Federal Reserve Board of New York may be cited as examples of the *sensitive* index numbers. With these may be contrasted the Economic Adviser's *general-purpose* index of wholesale prices based on 78 commodities in India; the Board of Trade Wholesale Price Index based on 200 items, the Economist index of 58 items and the Statist index of 45 items in Great Britain. In U.S.A. a wider range of quotations is in use: the

U.S. Bureau of Labour Statistics' Index of Wholesale Prices includes 450 items and Dun's Index 200 items.

Quality of Commodities.

Several varieties of a commodity may be on sale in the market. Generally speaking, the most popular or common variety is selected for index number construction. If it is found necessary to give special importance to a commodity, quotations for a few qualities of it may be included. For instance, in the Economic Adviser's index three varieties of sugar are considered: D 28, D 27 and E 27. If the marketing system in respect of a commodity is such that the prices for 'export' are different from those for 'internal' consumption' of the commodity, prices for both types of sales may be taken, as is the case with tea in the Economic Adviser's index for which two quotations are taken for tea for export and one quotation is taken for tea for internal consumption. Thus, for 78 commodities which enter the Economic Adviser's index, 130 quotations are used.

It is essential to specify the variety of the commodity selected and to stick to it throughout; otherwise, prices of different qualities of the same commodity may be quoted from the same place at different times. In such a case, the index number for a particular week may show a rise or fall over the previous week due only to a change in variety rather than to a fluctuation in the price of the variety used in the previous week. This would defeat the very purpose of the index number.

Grouping of Commodities.

The selected commodities may be classified into homogeneous groups, so that group index numbers of like commodities may also become available along with the all-commodities index. In the Economic Adviser's index now current, the 78 items are classified into 5 groups, *viz.*, (I) Food Articles, (II) Industrial Raw Materials, (III) Semi-Manufactures, (IV) Manufactures, (V) Miscellaneous.

Groups of commodities may be further sub-divided into sub-groups, to carry homogeneity a step further. In the Eco-

nomic Adviser's index, for instance, the Food Articles group is sub-divided into three sub-groups: (1) Cereals, (2) Pulses and (3) Others. The sub-group indices will afford an idea of the variation in the general price structure of the commodities falling in the sub-group.

Price Reporting Centres.

It is impossible to collect prices of the selected commodities from all the places in the country where they are bought and sold. A sample of markets will, therefore, have to be selected. The criterion will be to select those places where a given commodity is marketed in large quantities. In other words, representative trade centres will be chosen for each commodity. It is possible that a centre chosen as representative for one commodity may also be representative for several other commodities.

Price Reporting Agency.

In each selected reporting centre there may be a number of agencies to report price quotations, *e.g.*, business houses, chambers of commerce, trade associations, banks, news-correspondents, government agents, government price reporters, etc. A sample of these also will be taken for each reporting centre, unless the maker of index numbers chooses to bear the cost of setting up his own reporting agency. The criterion for sampling here will be the reliability and trustworthiness of the reporting agency. If there is reason to believe that a particular agency will be biased or motivated, it will have to be left out. As a check on the accuracy and reliability of price-reporting, it will be good to obtain quotations for the same commodity from more than one reporting-agency at the same reporting-centre. The quotations obtained from these different agencies may not be identical; they will, no doubt, be corroborative of the price trends provided the reporting has been unbiased.

Price Quotations.

Both the manner of quoting prices and the connotation of the term price should be clearly and unambiguously defined, intimated to the reporting agencies and strictly adhered to.

This will ensure collection of prices on uniform basis, which is an essential pre-requisite for comparison and for studying price variations.

In so far as the manner of reporting prices is concerned. prices should be quoted as so much money per unit for a commodity (*e.g.*, Rs. 5|8|- per maund) and not as so much units of commodity per unit of money (*e.g.*, 7 seers 4 chattaks per rupee). The former are called money prices and the latter quantity prices. Before 1907, prices in *Prices and Wages in India* published by the Government were expressed as seers and fractions of a seer per rupee. Use of such quantity prices needs proper care and caution. Money prices vary inversely with quantity prices and the percentage rise and fall also varies in the two notations. Thus, if rice sells at 4 seers per rupee and later changes to 2 seers per rupee, the quantity price would be said to show a fall from 100% to 50%. But the money price for rice would change from Rs. 10 per maund to Rs. 20 per maund. It would show a rise from 100% to 200%.

As regards the connotation of the term price, several considerations arise. For measuring changes in the general level of prices, the prices used in the index number should be wholesale (and, not retail). Wholesale prices are far more uniform over a region for the same day and are more sensitive to changes in the conditions of supply and demand than are retail prices. Retail prices lag behind wholesale prices both in their rise and fall and also fluctuate between narrower limits. Wholesale prices are, therefore, a better guide for disclosing movements of economic forces that affect or determine prices.

But, the term 'wholesale price' does not convey the same meaning to everybody. Is it the price for the naked commodity or should also include the incidental charges, such as, cost of container, weighment charges, sales tax, etc? Is it the price ex-godown, or free on the road|rail, or delivery-free at the purchaser's premises? Is it the opening rate in the market on the day the quotation is taken, or mid-day rate, or rate at a particular hour, or the closing rate? Is it the price at which, say, a wholesale dealer purchases from the farmer who

brings his card-load to the market for disposal, or the price at which the wholesale dealer sells off to other purchasers? Is it the 'controlled' price if the commodity is a controlled commodity, or the black-market rate at which illicit sales take place? These and similar other questions arise in determining the exact connotation of the term 'wholesale price.' Answers to them have to be found in the light of the purpose of the index number. Generally speaking, the wholesale prices used in index number construction are the prices for the naked commodity, ex-godown, at the close of the day and for sales made by wholesale dealer. If a commodity is controlled, controlled price is taken provided the controlled is generally effective.

Frequency of Quotations.

How often should be quotations be obtained In the case of weekly index numbers, one quotation per week for each commodity is generally considered to be sufficient. This quotation is usually obtained for a specified day of the week. In the Economic Adviser's index number of wholesale prices one-day-a-week prices on or about Friday are used. For a monthly index number, weekly quotations would be adequate, though more frequent quotations would improve the accuracy of measurement of price behaviour.

The importance of regular reporting of prices cannot be over-emphasized whatever be the frequency of their reporting. For, if price for a particular cmmodity is not reported in, say, a particular week, it may have to be omitted or estimated, or repeatd from the previous week, and to the extent the actual prices is different, the index for the week may be unreal. The extent of error introduced in the index on account of this factor may, however, be negligible in times of stable prices. The importance of regular reporting is, therefore, great when there are large and quite frequent changes in the market.

Averaging the Quotations.

If quotations for the selected variety for a commodity from a reporting centre are received on a daily or bi-weekly basis for constructing a weekly index, they may be averaged by

striking their simple arithmetic average. The average may then be used for constructing the weekly index. If only one-day-a-week quotation is obtained, the question of averaging does not arise for the weekly index.

The following table gives the annual average of the wholesale prices of certain commodities in Kanpur. It is proposed to illustrate the processes and methods involved in constructing general index numbers from the data given in this table.

Table 27. *Average Wholesale Prices of Certain Commodities in Kanpur,* 1928—1934.

Line	Commodities	Average Prices in Rs. per Maund						
		1928	1929	1930	1931	1932	1933	1934
1	Rice	7.3	7.7	5.8	4.1	4.3	4.1	3.7
2	Wheat	7.7	5.5	3.6	2.7	3.4	3.2	2.8
3	Linseed	7.0	8.0	6.5	4.2	3.5	3.4	3.6
4	Gur	6.5	7.3	6.2	4.2	3.5	3.1	4.1
5	Cotton	34.1	29.8	17.3	13.3	14.8	12.9	13.2
6	Tobacco	17.3	17.1	14.5	11.6	4.9	4.9	5.7

CHOICE OF BASE AND CALCULATION OF RELATIVES

After the prices have been collected, the next step in the construction of index numbers would be to reduce them to percentages or relatives. For doing so, an appropriate base in terms of which the prices shall be expressed as percentages should be selected. Two methods are available for this purpose:

I. Fixed Base.

II. Chain Base.

Fixed Base Method.

In the fixed base method either (i) the average price of some arbitrarily chosen *year* is taken as the base, or (ii) the average of the prices of a *period* of years is taken as the base. The base so chosen is adhered to for an indefinite time. In following the latter procedure, either the prices for three, five or even ten years may be averaged, or the prices of the entire period for which index numbers are to be constructed may be averaged. Averaging for the entire period is useful when the data are reviewed at the expiry of the period. Averaging for a few years or use of a particular year should be preferred if the data are to be made of a continuous character.

The year chosen as the base year should be a reasonably normal year. If an arbitrary year is chosen as the base, it may happen to be an abnormal year, for instance, a year of labour unrest, of war or of financial crisis. The index numbers calculated on an abnormal year will have to be qualified with a statement drawing attention to the abnormality of the base year. To avoid this, a base period is often chosen. Average of a period of years is representative, is less affected by chance variations and is most generally applicable in statistics. In India, however, the wholesale price indices and most cost of living indices use a single year as base year. Only in some cost of living indices is the average of a few years taken as the base. The base year of the Economic Adviser's wholesale price index is the year ending August 1939.

Computing Price Relatives.—The average price of the base chosen is taken as 100, and the price in each of the other years is expressed as a percentage of this amount. Thus,

$$\frac{\text{Price of the commodity for the current year}}{\text{Price of a commodity for the current year}} \times 100$$

will give the percentage (or price relative) for the current year. Let us choose 1928 as the base year to convert the data relating to other years, in table 27 above, into index numbers. The percentage price of rice in 1930 on the basis of 1928 is $\frac{\text{Rs. } 5.8}{\text{Rs. } 7.3} \times 100 = 79$. This price relative is the index number for rice for 1930 with 1928 as the base. All the relatives in table 28,

from line 1 to 6, have been computed in this manner with 1928 as the fixed base year.

Table 28. *Fixed-Base Relative Index Numbers of Wholesale Prices of Certain Commodities in Kanpur,* 1928—1934. (1928=100).

Line	Commodities	Percentages or Relatives, 1928=100						
		1928	1929	1930	1931	1932	1933	1934
1	Rice	100	105	79	56	59	56	51
2	Wheat	100	71	47	38	44	42	36
3	Linseed	100	114	93	60	50	49	51
4	Gur	100	112	95	65	54	48	63
5	Cotton	100	87	51	39	43	38	39
6	Tobacco	100	100	84	67	28	28	33
7	Total of Relatives	600	589	449	325	278	261	273
8	Average of Relatives	100	98	75	54	46	44	45
9	Median of Relatives	100	102	82	58	47	45	45
10	Geometric mean of Relatives	100	97	72	53	45	42	44

Chain Base Method.

In the fixed base method the base is fixed in the sense that the relatives for all the years are based on the prices of a single year (1928 in our example) or on the average prices of a period of years. Constrasted with it is the chain base or the shifting base method in which the relatives for each year are calculated upon the prices of the preceding year. Thus, the base year is not fixed, but changes from year to year. If, however, it is desired to convert these relatives to one common base, they are chained together.

Computing Link Relatives.—According to the chain base method, in our example, we would express the 1929 figures as percentages of those for 1928, and get index numbers for the

commodities for 1929 on 1928 as base; then, for 1930 we would express the 1930 figures as percentages of those for 1929 and obtain index numbers for 1930 on 1929 as base; and so on. Thus the percentage (or, link relative, as it is called in the case of chain base method) for rice for 1929 is $\frac{\text{Rs. } 7.7}{\text{Rs. } 7.3} \times 100 = 105$; for 1930 it is $\frac{\text{Rs. } 5.8}{\text{Rs. } 7.7} \times 100 = 75$; and so on. The link relatives in table 29 from line 1 to 6 are based on the preceding year, that is, the years are linked together.

Table 29. *Chain-Relative Index Numbers of Wholesale Prices of Certain Commodities in Kanpur,* 1928—1934.

Line	Commodities	Percentages or Relatives Based on Preceding Year						
		1928	1929	1930	1931	1932	1933	1934
1	Rice	100	105	75	71	105	95	90
2	Wheat	100	71	65	75	126	94	88
3	Linseed	100	114	81	65	83	97	106
4	Gur	100	112	85	68	83	88	132
5	Cotton	100	87	58	77	111	87	102
6	Tobacco	100	100	85	80	42	100	116
7	Total of Link Relatives	600	589	449	436	550	561	634
8	Average Link Relatives	100	98	75	73	92	94	106
9	Chain Relatives (1928=100)	100	98	74	54	49	46	49

TYPE OF AVERAGE TO BE USED.

The relatives arrived at by the fixed base method or the chain base method should be averaged to yield the required final index number. In theory, any form of average can be used for the purpose. In practice, however, we are to choose among (a) arithmetic average (b) median and (c) geometric mean. Line 8, 9 nd 10 of table 28 give, respectively, the arithmetic

average, median and geometric mean of price relatives computed according to fixed base method. These averages are the final index numbers of wholesale prices at Kanpur for different years with 1928 as the base.

Arithmetic Mean of Relatives in Fixed Base Method.

The arithmetic average has the advantage of being readily intelligible, but suffers from a bias which it is not easy to remove. It is too much affected by the extremes; it gives too much weight to increasing prices and too little to decreasing ones. The arithmetic average of relatives, as we shall see a little later, is not reversible. For these reasons, the arithmetic average does not reflect the typical movement of prices.

Median of Relatives in Fixed Base Method.

Median is the easiest to calculate, and enjoys an advantage over the arithmetic average in that it is but little affected by extreme items. Median is, therefore, very likely to be more typical of price movements than the mean. But it may not be possible to find an *actual* median, *e.g.*, in table 28 the last six medians had to be interpolated. Besides, median may be erratic when the number of items is small. Again, the median of relatives is not reversible. Therefore, median too is not a suitable form of average.

Geometric Mean of Relatives in Fixed Base Method.

The geometric mean is of value when items in a group are considered from the viewpoint of their relative differences rather than that of absolute differences. Therefore, it is reasonable to use it in computing index numbers where the items to be averaged are themselves relatives. It is indeed suitable for measuring the average ratio of change in prices for it gives equal importance to equal ratios of change. For instance, when geometric mean of relatives is taken, the effect of doubling of one price is perfectly counterbalanced by the halving of another. This is not the case with arithmetic average or median. Similarly, if the price of one commodity rises by 50% and that of another falls by 50%, the arithmetic average of

relatives will neither rise nor fall implying that there is no change in the price level, while, in fact, both the prices have changed. The geometric mean of the relatives would, in this case, show that there is a change in price. Table 30 illustrates these two ideas.

Table 30. *Fixed Base Index Number of X and Y Commodities;* (1941=100).

Commodities	1941 (base year)		1942		1943	
	Price	Relative	Price	Relative	Price	Relative
X	Rs. 5	100	Rs. 10	200	Rs. 7.5	150
Y	4	100	2	50	2	50
Total of Relatives		200		250		200
Geometric Mean of Relatives		100		100		87
Arithmetic Average of Relatives		100		125		100

The price of X commodity is double that of 1941 in 1942, and of Y commodity is half that of 1941 in 1942. The arithmetic average index number for 1942 on 1941 is 125 implying a 25% rise in general price level; but the geometric mean index number corrects this impression by showing that there is no change in the level of prices in 1942 as compared with 1941. Again in 1943, price of X commodity has risen 50% over that in 1941 and of Y has fallen 50% below that in 1941. The arithmetic average index number for 1943 on the base 1941 is 100 implying that there is no change in the level of prices, but the geometric mean corrects this impression by showing that the index number in 1943 falls to 86.7 or 87. From these examples it will be clear that the geometric mean, through its property of giving more weight to small items and less to large ones,

creates the effect of reducing the influence of upward movements in prices and increasing that of downward movements. This property is of great value in tracing the course of prices. Geometric mean has the further advantage that it makes possible the replacement of commodities which have ceased to be representative by those which have become representative, without affecting the balance of the index. Yet another advantage of this mean is that the index number calculated by using it is reversible, that is, change of base year can be made without affecting the proportionate change in the general index numbers. Geometric mean is, therefore, likely to be more typical of the changes in prices than are the arithmetic mean and the median. Its use in index number construction is growing, although arithmetic average has so far been largely used.

Chain Relatives.

In table 29, link relatives from line 1 to 6 have been calculated on the chain base method and totalled up in line 7. In line 8, average link relatives have been computed by dividing the totals in line 7 by 6, the number of commodities. These are the index numbers on the chain base principle, *i.e.*, the index for each successive year is based on the preceding year. If it is desired to base them all on 1928, they may be placed in a chain. This is done in line 9. These chain relatives are the index numbers for different years on the chain base method in respect to the year 1928. The process of chaining together the link relatives is as follows:

Average link relative for 1929 referred to 1928 is 98,
Average link relative for 1939 referred to 1929 is 75,
Average link relative for 1931 referred to 1930 is 73.

Then $\frac{98}{100} \times 75$ will give the chain relative index for 1930 on 1928, and $\frac{98}{100} \times \frac{75}{100} \times 73$ will give that for 1931 on 1928.

An easier method of changing the link relatives is shown

in the table below, wherein the required multiplication has been indicated in brackets:—

Chaining the Link Relatives

Year	Link Relatives, with preceding year as base	Chain Index, with 1928 as base
1928	100	100
1929	$98(\times\frac{100}{100}=)$	98
1930	$75(\times\frac{98}{100}=)$	74
1931	$73(\times\frac{74}{100}=)$	54
1932	$92(\times\frac{54}{100}=)$	49
1933	$94(\times\frac{49}{100}=)$	46
1934	$106(\times\frac{46}{100}=)$	49

Merits & Demerits of Chain Base.

The chain base method has two advantages. Firstly, it enables a direct comparison between one year and the year succeeding it. This is far more useful in business and commerce than the indirect comparison through a remote fixed base. Secondly, it makes possible the dropping of old items and inclusion of new ones since every pair of years is treated independently of every other. The necessity of adding to the list of items or substituting certain items in the existing list is frequently felt when computing a series of index numbers over a period of time, because of some commodities going out of use and new ones coming into fashion. If fixed base method were used, the entire index number series will have to be recalculated when commodity lists are altered. The chain base

method is, however, cumbrous and if an error is once committed it will be carried through the entire series in the chaining process as can be visualised from the above table. If, for instance, the product of $(75\times\frac{98}{100}$ is wrongly put at something different from the actual, all the chain indices from 1930 onwards will go wrong; the error will accumulate. Chain base is, therefore, very often given up in favour of the fixed base after the first few years of index number construction, that is, when the commodity list has been finalised.

Reversibility Test & Choice of Average.

An important property, which an index number should possess, is its reversibility, that is, it should satisfy what is called the Time Reversal Test. That average which caters for reversibility of an index number based on it has a superior claim for use in the construction of index numbers than the average which does not. Reversibility means that the index for the current year based upon the base year and the index for the base year based upon the current year should be reciprocal to each other. That is, the following equation should be satisfied.

$$P_{01}\times P_{10}=1\ ;\ \text{or}\ P_{01}=\frac{1}{P_{10}}$$

where, P_{01} stands for index for the current year on the base year omitting the factor 100, (*i.e.*, for price change in current year compared with base year), and P_{10} stands for index for the base year on the current year without the factor 100, (*i.e.*, for price change in base year compared with current year).

The arithmetic average of relatives is not reversible.— Line 3 of table 31 on next page gives in column (d) the arithmetic average of relatives for commodities A and B for 1931 on 1930 as the base, and in column (e) the arithmetic average of relatives for 1930 based on 1931. These are, respectively, 130.5 and 78.35 so that $P_{01}=\frac{130.5}{100}=1.305$, and $P_{10}=\frac{78.35}{100}=.7835$.

Now, $1.305\times.7835=1.02$, which is greater than 1. Therefore, the arithmetic average of relatives is not reversible.

Table 31. *Testing the Reversibility of Index Numbers.*

Line	Commodities	Price in 1930	Price in 1931	$\frac{\text{Year 1931}}{\text{Year 1930}} \times 100$	$\frac{\text{Year 1931}}{\text{Year 1930}} \times 100$
	(a)	(b)	(c)	(d)	(e)
		Rs.	Rs.		
1	A	10	15	150	66.7
2	B	45	50	111	90
3	Arithmetic average			130.5	78.35
4	Geometric mean			129	77.5

The Geometric mean of relatives is reversible.—Line 4 of table 31 gives in columns (d) and (e) the geometric average of relatives. These are 129 and 77·5 approximately, so that $P_{01} = \frac{129}{100} = 1{\cdot}29$, and $P_{10} = \frac{77{\cdot}5}{100} = {\cdot}775$. Their product is 1 (allowing for the adjustment of decimals). Therefore, the geometric mean of relatives is reversible.

There is yet another way of looking at the reversibility of index numbers. If a relative shows an average increase of, say, 25 per cent from the base year to the current year, then this should also be capable of being described as a decrease of 20 per cent from the current year to the base year. In table 31, using the arithmetic average we find that the level of prices in 1931 is 30·5 per cent higher over the prices of 1930. We might, therefore, say that the prices of 1930 are lower than those of 1931 by 23·4 per cent of the latter; that is, if the index for 1931 on 1930 is 130·5, it should be (100—23·4)=76·6 for 1930 on 1931, but actually it is 78·35 as shown in column (e), line 3. Therefore, the index number based on arithmetic average of relatives does not possess the property of reversibi-

lity. Using the geometric mean we find that the level of prices in 1931 is 29 per cent higher over the prices of 1930. We might, therefore, say that the prices of 1930 are lower than those of 1931 by 22·5 per cent of the latter; that is, if the index for 1931 on 1930 is 129, it should be (100—22·5)=77·5 for 1930 on 1931, which actually is the case as shown in column (e), line 4, table 31.

It is clear, then, that geometric mean stands the test of efficiency provided by the Time Reversal Test. As such, geometric mean can be said to perform satisfactorily its function of showing the required change in the phenomenon under study. Consequently, geometric mean is more suitable than the arithmetic mean or the median. Geometric mean can also be used with the chain base method. It is used by the *Board of Trade* in England in the construction of wholesale price indices on the chain base principle. Geometric mean is also used in the construction of 'Index Numbers of Wholesale Prices' which are being issued by the Economic Adviser to the Government of India since 1947.

SYSTEM OF WEIGHTING.

The 'Unweighted' Index Number.

So far, in the construction of index numbers we have used simple average and no *special* assumption has been made concerning weights. Distinction is very often made between weighted and unweighted index numbers, but it should be noted that every index number is weighted in some form. In computing the simple average of relatives each relative is counted once. Therefore, apparently, weights are unity in each case. A further thought would reveal that the change in the price of a commodity from one date to another is related to the commodity's price level on the first date. If in the base year the price of a commodity is unusually high, it will have an influence to correspond on the total, that is, it would have the same effect as actual weighting would. This can be easily verified by recalculating a given series of index numbers upon a few different bases by using the arithmetic average of relatives and then noticing that every fresh series differs, not

only in the absolute values of the index numbers, which is immaterial, but also in the relative values of the indices, which is significant. That this would be so is evident from the fact that index numbers where simple arithmetic average is used are not reversible. We may, therefore, conclude that even the so-called 'unweighted' index numbers are arbitrarily or haphazardly weighted, the arbitrary element being exercised by the choice or shifting of base year. We may also say that when simple average of relatives is used, change in the base year is equivalent to change in weights.

Implicit and Explicit Weighting.

In those index numbers which are termed "weighted' the weights are chosen according to some systematic plan. Weights may be implicit or explicit. Implicit weights relate to the selection of commodities themselves. If a particular commodity, or a commodity of the same general class, is included, say, 3 times in the list of prices, the weight given to the commodity is 3. For instance, 3 varieties of sugar may be included. Varying emphasis is, thus, given to the different items while selecting the commodities by the number of times a given commodity is included in the selection. Many of the so-called unweighted index numbers may, in fact, be indirectly or implicitly weighted. For instance, the *Bombay* Wholesale Price Index Numbers discontinued since 1944 were implicitly weighted.

In assigning explicit weights, weights proportional to the relative importance of different items are used. But, what considerations determine this relative importance ? This enquiry is essential because weights should either be appropriate or they should not be used at all. Now, in constructing an index to show general changes in prices, the weights assignable to wholesale prices may be several, for instance, the quantity of goods placed on the market, value of goods produced, values consumed, and so on. Different systems of weighting would yield different results. The difficulty, then, is: which of these or other similar criteria should be accepted as correct ? This difficulty is not easy to solve. Therefore, it appears that weights may better be ignored. This idea is strengthened by

the fact that weighted results are almost identical with the unweighted ones, if weights are chosen according to chance. According to Mitchell, the real problem before the maker of index numbers is whether he shall leave weighting to chance or seek to rationalize it. If chance weighting and rational weighting yield the same results, there is nothing much to choose between the two. The important point to remember is that weighting must be logical, that it should not be haphazard or irrational, that it should emphasize the real relative importance of the items entering the index.

While discussing the subject of weighted mean, Bowley observes: "The discussion of the proper weights to be used has occupied a space in statistical literature out of all proportion to its significance, for it may be said at once that no great importance need be attached to the special choice of weights; one of the most convenient facts of statistical theory is that, given certain conditions, the same result is obtained with sufficient closeness whatever logical system of weights is applied." He then illustrates that by using different weights, the resulting averages come out to be uniform in size. He also shows that the unweighted and the weighted means do not materially differ from each other. Bowley, however, is not opposed to weighting as such. He says at one place that paucity of data might make the inclusion of weights necessary and the popular desire for concrete measurements might make a fine show of weighting expedient. In the sphere of index numbers, he advocates weighted arithmetic average—arithmetic cross of Laspeyres and Paasche formulæ. Weighting seems necessary also because of the heterogeneous character of the series from which indices are computed. All prices entering an index are not of equal importance; all of them do not effect the country's pricing system in like manner. Most wholesale price indices in the U.S.A. are weighted. The Economic Adviser's wholesale price index in India is a weighted index. Weighting is, indeed, essential in constructing cost of living and business activity indices, as we shall see later.

---

[2] See Bowley, A. L., *Elements of Statistics*, 1920 ed., p. 86.

Methods of Weighting.

Two methods of explicit weighting may be distinguished: The Weighted Average of Relatives (Ratios) Method and the Weighted Aggregate of Actual Prices Method. The latter is known as the Aggregative Method also.

Weighted Average of Relatives.—According to this method *price relatives are weighted by values.* Values are obtained by multiplying quantities with their respective prices. The sum of the products of price relatives of the current year and values of the base year divided by the sum of the weights gives the weighted arithmetic average of relatives, which is the required index number for the current year. Symbolically,

$$\text{Index Number for Current Year} = \frac{\Sigma IV}{\Sigma V}$$

where, I stands for price relative and V for value (weight). Table 33 on page 233 demonstrates the working of this method. Weighted median and weighted geometric mean of the relative may also be computed.

Weighted Aggregative Method.—According to this method *prices themselves are weighted by quantities,* since total value is equal to price × quantity. The products of actual prices of the current year and quantities of the base year are summated. This sum is expressed as ratio in relation to a given base. This ratio is the index number for the current year. Symbolically,

$$\text{Index Number for Current Year} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100.$$

where p_1 stands for price in current year.
p_0 stands for price in base year,
q_0 stands for quantity in base year.

Table 32 on page 232 demonstrates the working of this method. Index numbers constructed by employing the Weighted Relatives and the Weighted Aggregative methods would be identical.

Summary.

The technique of construction of wholesale price index numbers may now be summarised as follows:—

(1) Select a reasonable number of representative commodities, and specify their varieties.

(2) Arrange for obtaining their regular wholesale prices from
  (i) either, standard trade journals,
  (ii) or, leading leaders of representative centres.

(3) Select a normal year, or average of years as the base.

(4) Reduce the average prices to percentages, *i.e.*, price relatives, on
  (i) either, the fixed base method, where
    (*a*) the fixed base may be a fixed year, or
    (*b*) it may be an average of a period of years,
  (ii) or, the chain base method.

(5) If the fixed base method is used, compute an average of relatives, using the arithmetic average, median or the geometric mean. Theoretically, the advantage lies with the geometric mean.
If the chain base method is used, chain together the link relatives.

(6) If weighting is necessary, compute
  (i) either, the weighted average of relatives,
  (*ii*) or, the weighted aggregate of actual prices.

Thus, we have discussed two important methods of constructing index numbers, viz., the Average of Relatives Method and the Weighted Aggregate of Actual Prices Method.

Limitations of Index Numbers.

A comparison of the unweighted index numbers calculated on the fixed base method in table 28 and of the unweighted index numbers calculated on the chain base method in table 29, and also of the weighted index numbers which can be calculated from the same figures would reveal that different methods yield different results, but all index numbers—without any excep-

tion—point to the same direction. Therefore, an index may be relied upon so far as the tendency shown by it is concerned without being trustworthy to the last digit. It is not the absolute value of an index number that matters. What matters is the general trend shown by it, or by a series of index numbers.

COST OF LIVING INDEX NUMBERS.

The methods of weighting discussed above are more particularly used in the construction of cost of living index numbers. These indices are designed to study the effect of changes in prices on the people as consumers, or, in other words, to study the average increase in the cost of maintaining the standard of living in a given year unchanged from that in the base year. General index numbers fail to afford us an exact idea regarding the effect of the change in the general price level on the cost of living of different classes of people, since a given change in the level of prices affects different classes of people differently. Therefore, to obtain a measure of the general movement of prices of those commodities which enter into the consumption of different classes of people, Cost of Living Index Numbers are compiled.

Difficulties in Constructing Cost of Living Indices.

Standard of living varies with income or occupation. Therefore, one single cost of living index will not be truly representative of people of different incomes. Consequently, index numbers should be compiled separately for different classes of people. But standard of living also varies with region or place in which people reside. This difficulty can be solved by compiling index numbers separately for different localities or different homogeneous zones. Again, same classes of people, at the same time, do not spend their income in exactly similar proportions on different objects. The best that can be done to obviate this difficulty is to collect a reasonable numbers of sufficiently accurate samples of family budgets from the same class of people to have a general idea of the propertions of expenditure on different objects by an *average* family. And, yet there is another difficulty that the same classes of people at different times spend their income in varying pro-

portions. A change in the nature and quantity of commodities consumed may arise from a change in tastes or fashions, or from an increased purchase of cheapening commodities and decreasing consumption of things becoming dearer. These factors, indeed, go a long way in explaining the change in the cost of living. But these changes cannot be taken stock of every year without incurring the huge expense of conducting fresh family budget enquiries. For this reason, it is assumed that the qualities and quantities of commodities consumed in the base year by a particular class of people remain the same for an indefinite period. These qualities and quantities, therefore, from the basis of the index number series. Another factor that causes a change in the standards of living is the change in the purchasing power of money. Cost of living index number confines itself to a measure of this factor alone. Further, people as consumers pay retail and not wholesale prices. Therefore, retail prices are taken into consideration in constructing cost of living indices. But retail prices vary from locality to locality. If cost of living index numbers are computed separately for different classes and different regions, this difficulty of variation in retail prices is also got over.

Construction of Cost of Living Index Numbers.

The first step that is taken for the construction of cost of living index number series is to decide the class of people—industrial workers, clerks, etc.,—for which the index numbers are to be compiled.

Next, a **sample budget enquiry** of the class concerned is conducted. The sample should cover a reasonably adequate number of families. The enquiry should be conducted during a period reasonably free from abnormalities of very high or very low prices. This budget enquiry would give precise information regarding (1) the nature, qualities and quantities of commodities consumed by the people, classified under the heads of food, clothing, rent, lighting and fuel, and miscellaneous groups, (2) the retail prices of the different commodities, (3) the proportion that the expenditure on each individual item of expenditure bears to the expenditure on the group to which it belongs, and (4) the proportion which expenditure on each group bears to the total expenditure.

The family budget enquiry forms the basis of the index number series. With it, the **selection of commodities** whose retail prices are to be regularly obtained becomes easy. It is important to note that a cost of living index number should include only those commodities under the food, clothing, etc., groups which are generally used by the class of people concerned, which are not subject to wide variations in quality nor to seasonal alterations in supply, and for which regular and comparable quotations of prices are obtainable.

Retail price quotations should be obtained from the localities in which the class of people concerned reside or from which they usually make their purchases.

The **sources of price quotations** may be either standard trade journals, or publications of government or municipalities, or typical businessmen in the locality concerned.

To convert the retail prices into index numbers, the prices or their relatives must, of necessity, be **weighted**, because the average consumer is not recompensed for a rise in the price of, say, cotton cloth by an equal fall in that of cement. Different objects of his consumption have different importance in his budget. They must be assigned their relative importance. For this purpose, one of the two systems of weighting already referred to in the foregoing pages may be applied: (1) The Aggregate Expenditure Method, also called the Aggregative Method, and (2) The Family Budget Method, also called the Method of Weighted Relatives.

Aggregate Expenditure or Aggregative Method.

This method is the same as the Weighted Aggregate of Actual Prices Method already discussed on page 227. Table 32 demonstrates the calculation of cost of living index number for the artisan class in the eastern U. P. by this method. (The figures in the table are imaginary). *Quantities* consumed in the base year have been taken as weights for the current year. The quantities consumed in the current year may be, and usually are, different from those consumed in the base year; but, for the reason already indicated (*viz.*, the enormous cost involved in

conducting a fresh budget enquiry every year), fixed weights, *i.e.*, quantities consumed in the base year, have been, and are, used as weights.

Table 32. *Construction of Cost of Living Index Number by the Aggregate Expenditure or Aggregative Method*

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Article	Quantities consumed in Base year (1925)	Unit	Price in base year (1925)	Price in current year (1941)	Aggregate Expenditure in base year [cl. 2×cl. 4]	Aggregate Expenditure in current year [cl. 2×cl. 5]
	q^0		p_0	p_1	$p_0\ q_0$	$p_1\ q_0$
			Rs.	Rs.	Rs.	Rs.
Rice	5 mds.	per md.	6	8	30	40
Bajra Jowar	5 mds.	,,	4	5	20	25
Wheat	1 md.	,,	5	10	5	10
Gram	1 md.	,,	3	6	3	6
Arhar	5 md.	,,	4	6	2	3
Other pulses	2 mds.	,,	3	4	6	8
Ghee	4 seers	per seer	1.25	2	5	8
Gur	2 mds.	per md.	2.50	5	5	10
Salt	12.5 seers	,,	4	5	1.25	1.6
Oil	24 seers	,,	20	25	12	15
Clothing	40 yds.	per yard	0.25	0.5	10	20
Firewood	10 yds.	per md.	0.50	0.8	5	8
Kerosene	1 Tin	per tin	4	6	4	6
House-rent	1 Hou.	per house	12	15	12	15
					$\Sigma p_0q_0=$ 120.25	$\Sigma p_1q_0=175.6$

$$\text{Index Number for the Current Year (1941)} = \frac{\Sigma\ p_1\ q_0}{\Sigma\ p_0\ q_0} \times 100$$

$$= \frac{175.6}{120.25} \times 100$$

$$= 146$$

Quantities consumed in any year (supposed to be typical) other than 1925 could also have been used as fixed weights.

Similarly, figures proportional to quantities consumed could also have been used in place of the actual figures.

Family Budget or Weighted Relatives Method.

This method is the same as the Weighted Average of Relatives Method already discussed on page 227. Table 33 demonstrates the calculation of cost of living index number for the same artisan class of the eastern U. P. by this method, using the same data. **Values** consumed in the base year have been used as weights for the current year. For the reason already indicated values in the current year are not used as weights, and fixed weights are employed.

Table 33. *Construction of Cost of Living Index Number by the Family Budget or Weighted Relatives Method.*

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Article	Unit	Price in base year (1925)	Price in current year (1941)	Price Relatives for current year	Weights (Values consumed in Base year)	Product of Price Relatives and weights [cl. 5×cl. 6]
		p_0	p_1	$p_1/p_0 \times 100 = I$	V	IV
		Rs.	Rs.		Rs.	
Rice	per md.	6	8	133.3	30	3999
Bajra Jowar	,,	4	5	125	20	2500
Wheat	,,	5	10	200	5	1000
Gram	,,	3	6	200	3	600
Arhar	,,	4	6	150	2	300
Other pulses	,,	3	4	133.3	6	799.80
Ghee	per seer	1.25	2	160	5	800
Gur	per md.	2.50	5	200	5	1000
Salt	,,	4	5	125	1.25	156.25
Oil	,,	20	25	125	12	1500
Clothing	per yard	0.25	0.5	200	10	2000
Firewood	per md.	0.50	0.8	160	5	800
Kerosene	per tin	4	6	150	4	600
House-rent	per house	12	15	125	12	1500
					$\Sigma V = 120.25$	$\Sigma IV = 17555.05$

$$\text{Index Number for the Current Year (1941)} = \frac{\Sigma IV}{\Sigma V}$$

$$= \frac{17555.05}{120.25}$$

$$= 146.$$

Values of any year other than 1925 could also have been used as weights. Figures proportional to actual values could also be, likewise, used. For instance, instead of using 30, 20, 5, etc. as weights we could have divided each of them by 5 and used 6, 4, 1, etc. as weights.

It will be seen that *the cost of living index numbers by both the aggregative and the family budget methods exactly agree.* Indeed, they should, if the weights relate to the same year. Family Budget method, or Weighted Average of Relatives method, is largely in use.

Group Indices.

In table 33, weighted average of all the articles has been *directly* computed. This process can be improved upon by (i) dividing the articles into food, clothing, etc. groups, (ii) weighting their price relatives by the proportion which expenditure on each article bears to the total expenditure on the group, (iii) obtaining weighted average index for each commodity group, (iv) weighting the group index numbers by the proportion which expenditure on each group bears to the total expenditure, and (v) obtaining the weighted average of the group index numbers. The final weighted average is the required cost of living index number. This process of double weighting is more scientific than that of computing a direct weighted average, and is in general use. In India, cost of living index numbers are computed by this method as will be seen in the relevant chapter.

Errors in Cost of Living Index Numbers.

Errors in cost of living index numbers arise from the following factors:—

(1) demarcating one class of people from another incorrectly:

(2) faulty selection of representative articles entering the cost of living of the class of people concerned;

(3) lack of representative character of price quotations—information relating to clothing, for instance, may relate, in part, to cloths rather than to clothes simply because fairly steady and reliable prices for the latter may not be available;

(4) faulty assignment of weights—weights may be deliberately manipulated;

(5) changes in the demand of various articles or their prices in the period under investigation. For instance, the budget enquiry for the construction of cost of living index numbers for industrial labour at Kanpur extended over two decades *viz.,* 1914 to 1934 during which period the price of firewood changed from 6½ annas per maund in 1914 to 13 annas in 1919 and to nearly 8 annas in 1930.

Unsatisfactory Character of Cost of Living Index Numbers.

Even if errors of the types enumerated above are not allowed to enter into the construction of cost of living index number, the index is not fully dependable. The reason is simple. The total monthly expenditure of two families of the same class may be equal, but the distribution of the expenditure over different objects may considerably vary according to the number of persons in the family, their age, sex, religion, caste and mode of living, and also according to the rise and fall in the prices of articles consumed. Index number does not consider the variations in the expenditure of the individuals. It considers only average or normal cases. Therefore, an index number pointing to the change in the average cost of living cannot be applied to every individual case in the class. It should be taken only as a guide to the direction, or the general trend, of the purchasing power of money for the class of people for whom it is meant.

Further, in constructing cost of living index we proceed on the assumption that the quantities or the values consumed in the base year or the typical year do not change, whereas they actually do. Standard of living of the same family undergoes change as time elapses and as prices, tastes, etc., change.

This fact is not taken into account by the index. The objection may be met by answering that the index number considers the increase in the cost of maintaining unchanged the base year standard of living. True, but is there any particular sanctity about the base-year standard of living? The base-year standard may not be adequate. Improvement in it may be necessary. The base year standard may be far too behind the current standard. Therefore, to make the index truly representative of current standard, family budgets should be collected regularly after an interval of a few years. In the light of every fresh inquiry, new weights and base should be adopted, and commodities, their qualities and quantities, all modified.

INDICES OF INDUSTRIAL PRODUCTION.

If it is desired to study the general change in the industrial activity in a country over a period of time, evidence of this change may be found in changes in the output of the various industries of the country. The first step, therefore, will be to collect information relating to the production of different groups of industries. Informations may, for instance, be had for the following groups of industries:

1. Mining—coal, iron ore, petroleum, gold, manganese.
2. Metallurgical—steel work, rolling mills, foundries etc.
3. Mechanical—locomotives, ship-building, railway rolling stock etc.
4. Textile—cotton, woollen, jute, silk etc.
5. Industries usually subject to excise duties—distilling of alcoholic beverages, brewing, sugar, matches and tobacco etc.
6. Other important industries—chemical, cement, glassware, flour-milling, oil-crushing etc.

The statistics of production of these industries may be available on monthly or annual basis. The production figures of the base year may be put down at 100 and those of the subsequent month or years expressed as percentages of the

base year. The resulting production relatives are then multiplied by the weights. Weights may be in proportion to the relative importance of the industries in the country, the importance being determined by the capital invested, or net output (that is, added value of manufactures), or some other criterion. The weighted average—arithmetic or geometric—of the relatives gives the index number of industrial production for the country. The index may be constructed for gross output or net output, as desired.

To attempt a study of the changes in the business conditions of a country, it would be necessary to collect far more comprehensive data than are required for computing indices of industrial production. Professor Pigou selected the following series for a study of the changes in business conditions of England:

1. Unemployment percentage.
2. Consumption of pig-iron.
3. Prices in England.
4. Rates of discount on three months' bills.
5. Volume of manufactured goods.
6. Agricultural production.
7. Yield per acre of nine principal crops.
8. Index of production from mines.
9. Clearings of London Clearing House.
10. Increase of bank credit.
11. Credits outstanding.
12. Annual increase in the aggregate money wage.
13. Rate of real wages.
14. General aggregate consumption.
15. Proportion of reserve to liabilities of the Bank of England.

These quantities may be converted into relatives referred to a base year. From these relatives a weighted average may

be obtained. This weighted average shall be the Index Number of Business Conditions. It will afford a general idea of the average change in the business conditions of the country and serve as Economic Barometer or Forecaster of changes in business conditions through periods of depression, recovery, prosperity and crisis. Business conditions are never stationary. They do change, but the change, it has been found by experience in industrial countries of the west, particularly the U.S.A., is not fortuitous. These changes are also not regular and periodic. Business in general passes through well-defined major and minor changes. Accordingly, it is possible to study their order, to measure their present conditions and to give a forecast of likely position in the future. Both in England and in the U.S.A. interest in this subject is growing.

BASE SHIFTING, SPLICING AND DEFLATING

Base shifting.

Not infrequently it becomes necessary to shift the base of a series of index numbers from one period to another. For instance, if a series of indices based on the year 1914 is to be compared with a similar series whose base year in 1939, then in the interest of valid comparison it will be necessary to convert the former series into a new one with 1939 as the base year. That is, the base of the former series will be shifted to a new one.

From what has been said above in regard to reversibility of index numbers, it is clear that index numbers based on geometric mean of relatives can be shifted from one base to another without error by what may be called the 'short-cut' method, illustrated in the example on pages 223-224. But it is not possible to shift the base *without* error by using the same method when arithmetic average has been used in averaging the relatives. The 'short-cut' method is, of course, not possible to apply when median is used in averaging the relatives.

The general principle involved in base shifting is that, since index numbers in a series are ratios, they may be divided or multiplied throughout by any desired constant. The table below gives chain indices with 1928 as the base in col. (2). If

it is desired to shift their base to 1930, that may be achieved by dividing the series by the index for that year. The required arithmetic is indicated in brackets, and the new series in col. (3).

Shifting the Base

(1) Year	(2) Chain index (Base: 1928)	(3) Chain index (Base: 1930).
1928	100 $(\div \frac{74}{100} =)$	135
1929	98 $(\div \frac{74}{100} =)$	132
1930	74 $(\div \frac{74}{100} =)$	100
1931	54 $(\div \frac{74}{100} =)$	73
1932	49 $(\div \frac{74}{100} =)$	66
1933	46 $(\div \frac{74}{100} =)$	62
1934	49 $(\div \frac{74}{100} =)$	66

The above method can as well be used in the case of index numbers computed on fixed base method.

Another method for base shifting is also available, viz., re-computing the relatives of each individual item on the new base and averaging their total, that is, *reconstructing* the entire series. This method should be used for shifting the base when arithmetic average or median has been employed in averaging the relatives of a series, while both this and the short-cut methods will yield identical results when geometric mean has been used.

Splicing two Index Numbers.

Suppose an index number series A, computed with 1900 as base, is discontinued in 1940; and series B is computed with 1940 as the base, and it is desired to build a continuous series from 1900 onwards, that is, to make B continuous with A. This can be done by splicing B to A on the basis of the ratio between A and B in the year of overlap, *viz.*, 1940, as will be seen from the following table:

Splicing two index numbers

Year	A	B	B spliced to A
1900	100		
..	..		
..	..		
1939	160		
1940	150	100 (×150/100)	150·0
1941		95 (× 150/100)	142·5
1942		98 (× 150/100)	147·0
1943		96 (× 150/100)	144·0

Dividing A by the ratio 150/100 would make A continuous with B.

Deflating.

Deflating stands for making allowance for the effect of changing price levels. Thus, bank clearings over a period of time may be deflated by dividing each year's clearings by the corresponding general index number. Likewise, a wage series may be deflated by dividing each item by the cost of living index. By doing so, real wages will result, which may be converted into index numbers if desired. It is obvious that in deflating only that index number should be used which is appropriate to the given case.

The following table illustrates the process of deflating *per capita* national income by the cost of living index:

Deflating Per Capita Income.

Year	*Per Capita* income Rs.	Cost of living index (Base: 1939-40)	Real income *per capita* Rs.
1939-40	67	100	67·0
1940-41	70	105	66·6
1941-42	78	117	66·6
..	..	..	..
1947-48	160	258	62·0

FISHER'S "IDEAL" FORMULA

The fixed-weight aggregative method is very popular. It is used in constructing the Bureau of Labour wholesale price index in U.S.A. In the theory of index numbers, one point that has received considerable attention is the cross of base year and current year weights. If quantities for all the years for which it is desired to calculate the index numbers are available, the weights may be made to vary from year to year, quantities for different years being used as weights for their respective years. Of the several formulæ that have been suggested for this purpose, that given by Fisher may be considered in some detail.

Professor Irving Fisher[4] after an elaborate examination of 134 possible formulæ concluded that a scheme of cross-weighting should be used, and gave a Crossed Weight Formula, which is also named as Fisher's "Ideal" Formula. It is :

$$\sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

This formula requires four sets of aggregates, *viz.*,

(1) $\Sigma p_1 q_0$: Current year price $\times$ base year quantity,
(2) $\Sigma p_1 q_1$: Current year price $\times$ current year quantity,

[4] See Fisher, Irving, *Making of Index Numbers*, 1922.

(3) $\Sigma p_0 q_0$: Base year price $\times$ base year quantity,
(4) $\Sigma p_0 q_1$: Base year price $\times$ current year quantity.

The first aggregate is divided by the third, and the second by the fourth. The two resulting relatives are multiplied together and square root of the product is extracted. Fisher calls this formula as "ideal", since it neutralizes the types of bias which are found in measuring prices and quantities. The system of weighting has been so designed in the formula that the resulting index satisfies two basic tests, *viz.*, Time Reversal Test and Factor Reversal Test.

Time Reversal Test.—It has already been indicated in connection with the "Reversibility Test and Choice of Average" on page 222 what time reversal test implies. According to Fisher this test may be described as follows :

> "The test is that the formula for calculating an index number should be such that it will give the same ratio between one point of comparison and the other, *no matter which of the two is taken as base.*
>
> "Or, putting it another way, the index number reckoned forward should be the reciprocal of that reckoned backward."[5]

This implies that the following equation should be satisfied:

$$P_{01} \times P_{10} = 1$$

This, again, means that if an index shows that between 1938 and 1942 prices doubled, then it should also show that the level of prices in 1938 was one-half of that in 1942 when measured from 1942.

Factor Reversal Test.—A second fundamental test by means of which good index numbers can be detected is the factor reversal test. Regarding this test Fisher says:

> "Just as our formula should permit the interchange of the two times without giving inconsistent results, so it ought to permit interchanging the prices and quantities without giving inconsistent results—*i.e.*, the two results multiplied together should give the true value ratio."[6]

[5] Fisher, Irving, *Op. Cit.*, p. 64.
[6] Fisher, Irving, *Op. Cit.*, p. 72.

This implies that the following equation should hold good:

$$P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

where, P_{01} stands for the price change for the current year on the base year, Q_{01} for the quantity change for the current year on the base year, p_1q_1 for the total value (price $\times$ quantity) in the current year, p_0q_0 for the total value in the base year, and $\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$ for the ratio of the total value in the current year over the total value in the base year.

The following example will illustrate that Fisher's "Ideal" Formula satisfies the two tests mentioned above. The under-mentioned data are given :

Commodity	Base Yr. Price	Base Yr. Quantity	Current Yr. Price	Current Yr. Quantity
	p_0	q_0	p_1	q_1
A	4	20	6	18
B	5	15	6	12
C	2	30	3	30
D	1	50	1	60
E	3	25	5	28

The four sets of aggregates required for computing Fisher's index may be computed as under from the given data:

Commodity	p_0q_0	p_1q_0	p_1q_1	p_0q_1
A	80	120	108	72
B	75	90	72	60
C	60	90	90	60
D	50	50	60	60
E	75	125	140	84
Total ..	340	475	470	336

Now, $P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} = \sqrt{\frac{475}{340} \times \frac{470}{336}} = \sqrt{\frac{22325}{11424}}$

And, $P_{10} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}} = \sqrt{\frac{336}{470} \times \frac{340}{475}} = \sqrt{\frac{11424}{22325}}$

Therefore, $P_{01}.\ P_{10}=\sqrt{\dfrac{22325}{11424}\times\dfrac{11424}{22325}}$

$$=1$$

The Time Reversal Test is thus satisfied. As for the Factor Reversal Test, we already know P_{01}, $\Sigma p_1 q_1$ and $\Sigma p_0 q_0$.

$$Q_{01}=\sqrt{\frac{\Sigma p_0\ q_1}{\Sigma p_0\ q_0}\times\frac{\Sigma p_1\ q_1}{\Sigma p_1\ q_0}}=\sqrt{\frac{336}{340}\times\frac{470}{475}}$$

$$=\sqrt{\frac{7896}{8075}}$$

Therefore, $P_{01}.\ Q_{01}=\sqrt{\dfrac{22325}{11424}\times\dfrac{7896}{8075}}$

$$=1{\cdot}382$$

And, $\dfrac{\Sigma p_1 q_1}{\Sigma p_0 q_0}=\dfrac{470}{340}=1{\cdot}382$

Thus, $P_{01}.\ Q_{01}=\dfrac{\Sigma p_1\ q_1}{\Sigma p_0\ q_0}$

The Factor Reversal Test is also, therefore, satisfied.

Fisher's "Ideal" formula not only satisfies the two tests of efficiency, but is also easy to calculate from the practical point of view. Therefore, of the 134 possible formulæ which Fisher analyzed, the "Ideal" is ideal. But this formula requires statistics of quantities for the base year as well as the current year. These statistics are generally not available for every current year. They may be available at each successive census of production, if such censuses are taken in a country. Therefore, the choice has to lie with the use of fixed weights, *i.e.*, quantities of the base year or the year supposed to by typical. Hence the general popularity of the fixed-weight aggregative.

PROBLEM OF PERFECT FORMULA FOR INDEX NUMBER CONSTRUCTION

The objective in constructing a series of index numbers is to find numbers, year by year, which would move in the same ratio as the phenomenon which is attempted to be measured through them. Attempts have, therefore, been made by statisticians from time to time to evolve formulae that would measure the movements in a given phenomenon as accurately as possible. It may, however, be stated that there is such thing as a 100% perfect formula, that no one formula is good enough for measuring all phenomena and that no one formula for measuring a phenomenon may be suitable for all times. The choice of index number formula, therefore, depends on the objects of the index number series. It also depends on the availability of adequate and reliable data. A formula may appear to be sound in theory, but may not be practicable for lack of requisite data.

Developments in Index Number Theory.

Developments in the theory of index numbers have occurred in association with price upheavals in different countries. The first among the modern writers was Jevons, who attempted a measurement of the effects of gold imports on price levels after 1848. He thought that variations in price relatives were variations in percentages and as such they followed a geometric law. He, therefore, used geometric mean for his famous index number in 1863. He handled the issue of weighting by omitting those 3 items from his series of 39 commodities which showed exceptionally large fluctuations. Marshall and Edgeworth wrote in the 1880's, when the problem of bimetallism and its relation to price levels was an issue of great significance in England and the United States. Edgeworth was the champion of the median, for he thought that the median was less affected by irregular and unusual relatives than were other averages. Mitchell introduced the idea of weighting the wholesale price index number of the U. S. Bureau of Labour Statistics in 1915. He advocated the method of fixed-weight aggregative. Fisher advocated a geometric cross of current year and base year weights in 1920 and after a discussion of 134 different formulæ

characterised one of them as "Ideal". Bowley, Walsh and Pigou developed crossed-weight formulæ.

Important Formulae.

Of the various formulæ, the following important ones may be mentioned:—

$$P_{01}=\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \quad \ldots\ldots (1)$$

$$P_{01}=\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \quad \ldots\ldots (2)$$

$$P_{01}=\tfrac{1}{2}\left[\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}+\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\right] \quad \ldots\ldots (3)$$

$$P_{01}=\sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}\times\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \quad \ldots\ldots (4)$$

$$P_{01}=\frac{\Sigma(q_0+q_1)p_1}{\Sigma(q_0+q_1)p_0} \quad \ldots\ldots (5)$$

$$P_{01}=\frac{\Sigma\sqrt{q_0 . q_1}p_1}{\Sigma\sqrt{q_0 . q_1}p_0} \quad \ldots\ldots (6)$$

$$P_{01}=\frac{\Sigma p_1 q_a}{\Sigma p_0 q_a} \quad \ldots\ldots (7)$$

The above formulæ are in terms of prices. They may be converted into formulæ for quantity index numbers by interchanging p and q.

In the notation used in the above formulæ,

P_{01}=price index number for the current year,

p_1=current year price,

q_1=current year quantity,

p_0=base year price,

q_0=base year quantity,

q^a=quantity for the year to which weights relate.

Of the above formulæ, (1) and (2) are generally regarded as very good; (3), (4), (5) and (6) as superlative; and, (7) as fair. This is from the view-point of the theory of index numbers.

The central four formulæ depend upon the whole data of both the periods of a comparison. Of these four formulæ, (3) and (5) are based on weighted arithmetic average and (4) and (6) on the weighted geometric mean.

Formula (1) is known as Laspeyres formula and formula (2) as Paasche formula. Formula (1) is base-weight aggregative, and formula (2) current-year-weight aggregative. Both these formulæ are recognised as good and of equal status as measures of price and quantity change. Some statisticians have, therefore, suggested a compromise between the two, the compromise being the mean of the two. Two such means have been suggested: arithmetic and geometric. Formula (3) is an arithmetic cross of Laspeyres and Paasche formulæ, while formula (4) is the geometric cross of Laspeyres and Paasche formulæ. Arithmetic cross, that is, formula (3), was recommended by Drobisch and Bowley. The geometric cross, that is, formula (4), is generally known as the Ideal formula, a name given to it by Fisher in 1920. It was also advocated by Pigou. A further advance on the above formulæ is that of cross-weighting. Formula (5) is arithmetically crossed-weight aggregative and formula (6) geometric crossed-weight aggregative. The former is often called the Marshall-Edgeworth formula because of its advocacy by them. The latter is regarded as one of the best formulæ in theory by Walsh.

Most Frequent Formula.

The most popular formula, however, is the fixed-weight aggregative, formula no. (7). It is used more frequently than any other formula in the construction of published index numbers of prices and quantities today. The fixed-weight in this formula measures in some way the importance of the prices. It should be clearly noted that in this formula, unlike the other 6 formulæ, the weights need not necessarily refer to the base year or the current year. They may relate to any year, or period of years, considered reasonable or appropriate by the

maker of index numbers. Collection of data relating to weights for the current year or the base year or both the years is not always feasible in every country. Since the fixed-weight aggregative provides a solution to this difficulty, it has naturally acquired general popularity. Another advantage of this formula is that it satisfies what is called the 'circular test'. This test implies that index numbers derived from closely related data should be consistent for all dual comparisons. For instance, if the index of the second year is twice that of the first year and the index of the third year is twice that of the second year, then the index of the third year should be four times that of the first year. Fisher's 'Ideal' formula, no doubt, satisfies the time reversal and factor reversal tests. It does not, however, satisfy the circular test. The circular test can, in fact, be satisfied only by unweighted or constant-weighted aggregatives or geometric averages of relatives. The fixed weight aggregative was first advocated in England in the first quarter of the nineteenth century. It was also suggested by Mitchell in U.S.A. in 1915 when he recommended its adoption by the U. S. Bureau of Labour for their wholesale price index number, which was then being compiled on the basis of a simple unweighted average of relatives.

The fact that the fixed weights in the weighted aggregative may not relate to the base or current year but to a different year, has also been criticised as a weakness of this formula. An index number provides a comparison between the base year and the current year. Then, why should not the weights refer to both or atleast either of the two years? This is a question that has been asked. In this question is summed up the weakness of the fixed-weight aggregative. This weakness has, however, been sought to be removed by (a) changing the weights oftener and (b) changing the base year also oftener, so as to bring both the weights and the base year as close to the period of study as possible.

Place of Chain Base.

It has also been held by some writers on index numbers that index numbers should satisfy the 'homogeneity test', that is, the data for two years which are compared through index numbers should be as homogeneous as possible. Homogeneity,

however, is at its maximum between two adjacent years. For that reason, the chain base method has been advocated in substitution of the fixed base method. By introducing homogeneity, the chain base method is claimed to improve the accuracy of the measurement of the change in the given phenomenon between two years. The link relatives may, of course, be finally chained together to constitute a time series.

The chain base method has other advantages too. Since every pair of years is treated independently of every other, new data and new weights may be taken in without affecting comparibility and without requiring the recalculation of index numbers which becomes necessary in the case of fixed base method. Marshall and Edgeworth advocated chain index numbers as the best means of making long time comparisons. Mitchell, however, objected to the chain index. He thought that in the chain index errors were likely to accumulate, that the chain tended to drift away from the fixed base and that the interpretation of the final results was simpler in the fixed base system. He, however, did not condemn the chain system utterly; he recommended that indices could be constructed on both the fixed weights and the chain base methods and attention then invited to differences between the two. Fisher thinks that chain system is of little or no real use, and that the fixed base is simpler to conceive and to calculate. The chain base, however, has the advantage of homogeneity of data between two adjacent years and that is thought, by some writers on index numbers, to be a sufficient reason to prefer it to the fixed base method.

Uses of Index Numbers.

Index numbers reflect the movement of some quantity to which they relate. Their peculiar character is that they exhibit the relative rather than the absolute aspect of such movement. Index numbers are not restricted to the price phenomenon alone. Any phenomenon which is stretched over a period of time and expressed numerically may be presented through them. They may, for instance, be designed to show changes in wages, values of exports and imports, prices of securities, production of certain manufactures, circulation of notes etc. Different index numbers serve different purposes.

General price index numbers measure general changes in prices and through them the value of money. If general prices, as indicated by the price index number of a certain year, double, the purchasing power of money for that year would be halved. Prices can be brought down or controlled through either the regulation of the supply of money and credit, or the regulation of production, or both. In any case, index numbers will provide an apparatus to study the fluctuation of general prices and a standard for keeping them steady in the interest of consumer, trade and public finance.

General price index numbers make possible a study of the movements of prices in different countries and of the fact whether they are fairly stable.

Cost of living index numbers indicate through their movement whether real wages are rising or falling, money wages remaining unchanged. They can be used to grant bonus to employees to meet the increased cost of living. Claims of labour for increase in wages, if they turn upon rising cost of living, can be sifted on the basis of cost of living indices.

Indices of industrial activity can be utilized to study the progress of general industrialisation of a country and the effect of tariff on the development of particular industries. When an industrial plan is being implemented, such indices are of immense use in judging the results of the policy adopted.

Indices of business conditions measure the change in the general economic activity of a country and afford an approximate idea of the fluctuations in the real national income of the country. They can be made to forecast economic events.

Investment index numbers are of great help to those interested in the stock market. Indices of imports and exports give an idea of the fluctuations in the foreign trade of a country.

It should, however, be noted that index numbers that are good for one purpose may not be useful for another. For instance, general index number indicating a rise in general price level is not a good guide for movement in cost of living, or, cost of living index numbers for industrial workers are of no use for the upper classes.

EXERCISES

(1) What considerations would weigh with you in regard to selection of commodities and the base year while constructing a wholesale price index number ?

(2) Distinguish between the Fixed Base and the Chain Base Methods of constructing index numbers and discuss their relative merits.

(3) Which average, do you think, is appropriate to use in averaging the price relatives to arrive at the final index number ?

(4) 'Index numbers are economic barometers.' Explain this statement, and mention what precautions should be taken in making use of any published index numbers.

(B. Com., Alld., 1952).

Show, with the help of an example, how you would convert the index number from one basic period to another.

(B. Com., Agra, 1940).

(5) Describe with illustration the weighted index number of wholesale prices, and show its importance.

(B. Com., Nagpur, 1942).

(6) Examine the claims of (*a*) Geometric average, and (*b*) Chain base method in the technique of index numbers construction. Illustrate your answer with examples.

(B. Com., Delhi, 1953).

(7) 'Averages linked with percentages constitute the whole basis upon which is raised the superstructure of a simple device—that of Index Numbers—of comparing factors which are not *directly* comparable.' Elucidate.

(B. Com., Alld., 1948).

(8) Define an index number. Explain the role of weights in the construction of an index of the general price level.

(M.A., Raj., 1950).

(9) Discuss the problem of obtaining a perfect formula for an index number of prices. Explain fully what is meant by the reversibility of an index number.

(M.A., Patna, 1940).

(10) "......the method of index numbers is at once applicable to the disentanglement of that which is common to

the whole group from those variations which are special to individual items." Elucidate.

(M. Com., Raj., 1952).

(11) 'The real problem for the maker of index numbers is whether he shall leave weighting to chance or seek to rationalize it.' (*Mitchell*).

Distinguish clearly between 'chance weighting' and 'rational weighting' and suggest a solution of the above problem. Also discuss whether Fisher's Ideal Formula offers a rational system of weighting.

(M. Com., Alld., 1944 & 1951).

(12) Index numbers seek to set aside the irregularity of individual instances and replace it by the regularity of the big numbers.—Comment.

(M. Com., Alld., 1947).
(M.A., Pb., Apr., 1952).

(13) 'The discussion of the proper weights to be used has occupied a space in statistical literature out of all proportion to its significance, for it may be said atonce that no great importance need be attached to the special choice of weights; one of the most convenient facts of statistical theory is that, given certain conditions, the same result is obtained with sufficient closeness, whatever logical system of weights is applied.' (*Bowley.*)

Discuss the above statement.

(M. Com., Alld., 1948).

(14) Describe briefly the method you will adopt for the compilation of "Cost of Living Index Numbers" for the working classes in an industrial area.

(B. Com., Hons., Andhra, 1944).

(15) Show, giving suitable examples, the importance of the use of Index Numbers in interpreting economic effects.

(B. Com., Alld., 1946).

(16) Explain Fisher's "Ideal" Method of weighting index numbers and state the difficulties that are to be faced in using it.

(17) What points would you take into consideration in choosing the base and determining the weights in the preparation of the Cost of Living index numbers?

(B. Com., Agra, 1943).

(18) It is desired to find the difference in the cost of living in the years 1939 and 1953 in the case of (*i*) clerks and (*ii*) industrial labourers in a big industrial town.

Explain fully the necessary procedure to be adopted.

(19) What are the main sources of errors in cost of living index numbers ? How can these errors be avoided ?

(B. Com., Alld., 1938).

(20) Explain the uses of Index Numbers. Describe the procedure followed in the preparation of general and cost of living index numbers.

(B. Com., Agra, 1942).

(21) Explain the use of Index Number with the help of the following table, which gives the average annual wholesale price of jute in Calcutta in rupees per bale of 400 lbs. for the period 1914 to 1930:—

Year	Rupees	Year	Rupees
1914	78	1922	88
1915	54	1923	78
1916	67	1924	76
1917	56	1925	112
1918	72	1926	99
1919	102	1927	76
1920	98	1928	75
1921	94	1929	71
		1930	50

(B. Com., Cal., 1934).

(22) Given the following data, what index numbers would you use for purposes of comparison ? Give reasons.

Year	Rice		Wheat		Jowar	
	Price	Quantity	Price	Quantity	Price	Quantity
1927	9·3	100	6·4	11	5·1	5
1934	4·5	90	3·7	10	2·7	3

Prices and quantities are given in arbitrary units.

(M.A., Cal., 1937).

(23) Explain what is meant by the Factor Reversal Test. Construct with the help of data given below Fisher's Ideal index, and show how it satisfies the factor reversal test.

	Estimated total produce in thousand tons in district Saran		Harvest Price per maund in district Saran	
	1931-32	1932-33	1931-32	1932-33
			Rs. a.	Rs. a.
Winter Rice ..	71	26	3 8	3 2
Barley ..	107	83	2 0	1 14
Maize ..	62	48	2 9	1 12

(M.A., Patna, 1942).

(24) Which average would you use in computing the Price Index Number from the following data for 1943 on the basis of 1942 ? Give your reasons.

Commodity	Unit	1942	1943
Wheat	per maund	Rs. 8-8-0	Rs. 17-0-0
Ghee	per maund	Rs. 50-0-0	Rs. 75-0-0
Firewood	per maund	Rs. 1-0-0	Rs. 00-8-0
Sugar	per seer	Rs. 0-9-0	Rs. 00-4-6
Cloth	per yard	Rs. 0-5-0	Rs. 00-2-6

(Figures in the above table are arbitrary).

(25) Compute the weighted average index for 1948 from the hypothetical price data given below. Use 1947 as base.

Commodity	Price in 1947 (Per unit)			Price in 1948 (Per unit)		
	Rs.	a.	p.	Rs.	a	p.
A	4	0	0	6	0	0
B	10	0	0	12	0	0
C	0	8	0	0	8	0
D	0	12	0	1	0	0
E	1	4	0	0	15	0

(M.A., Alld., 1950).

(26) Using 1939 as the base year, construct cost of living index number for 1940 from the following table:—

Article	Weights	Price in 1939		Price in 1940		Unit
		Rs.	As.	Rs.	As.	
Rice	8	8	0	10	0	per maund
Wheat	15	5	0	8	0	,, ,,
Pulses	6	6	0	7	0	,, ,,
Sugar	4	0	4	0	6	per seer
Ghee	5	1	4	2	0	,, ,,
Cloth	10	0	8	0	10	per yard
Firewood	5	1	4	1	14	per maund
Cigarettes	3	0	5	0	7	per packet
Paper	1	0	3	0	5	per quire
Kerosene oil	3	0	4	0	4	per bottle

(M.A., Alld., 1951).

(27) (*a*) Discuss briefly the problems involved in the construction of index numbers of prices.

(*b*) Calculate three sets of index numbers with the three quinquennial averages (i) 1909-10 to 1913-14, (ii) 1914-15 to 1918-19, and (iii) 1919-20 to 1923-24 as the representative bases from the following data:—

Year	Value of imports in Crores of Rs.
Average 1909—10 to 1913—14	145·9
Average 1914—15 to 1918—19	147·8
Average 1919—20 to 1923—24	254·0
1924—25	246·6
1925—26	226·2
1926—27	231·2
1927—28	249·8
1928—29	253·3
1929—30	240·8
1930—31	164·8
1931—32	126·4
1932—33	132·6
1933—34	115·4

(B. Com., Hons., Andhra, 1943).

(28) Discuss how you would compare the average ratio of prices in 1937 with that in 1927, commenting on the relative advantages and disadvantages of alternative methods which might be used for such comparison:

Commodity	Index Numbers (July 1914=100) 1927	1937
Jute raw ..	93	56
Jute manufacturers ..	146	67
Cotton raw ..	167	89
Cotton manufacturers ..	159	117
Wool and silk ..	126	126

(29) What is Chain Base Method ? Describe it in connection with the construction of index numbers from the following data.

Year		A	B	C	D	E
		Commodity Index Numbers				
1938	..	98	78	82	96	96
1939	..	100	82	78	100	100
1940	..	112	82	78	102	104
1941	..	110	84	84	98	98
1942	..	110	84	85	98	100
1943	..	120	90	90	100	100

(30) Following are the group index numbers and the group weights of an average working class family's budget. Construct the cost of living index number by assigning the given weights.

Group		Index Number for January 1943	Weights
Food	..	152	48
Fuel and Lighting	..	110	6
Clothing	..	130	8
House rent	..	100	12
Miscellaneous	..	90	15

(31) From the data given below calculate the cost of living index number for the current year by the Aggregate Expenditure and Family Budget Methods, separately.

Article	Quantity consumed in base year	Unit	Price in base year Rs.	Price in current year Rs.
Rice ..	5 maunds	maund	6	8
Millets ..	5 maunds	maund	4	5
Wheat ..	1 maund	,,	5	10
Gram ..	1 maund	,,	3	6
Arhar ..	½ maund	,,	4	6
Other pulses	2 maunds	,,	3	4
Ghee ..	4 seers	seer	1·25	2
Gur ..	2 maunds	maund	2·50	5
Salt ..	12½ seers	,,	4	5
Oil ..	24 seers	,,	20	25
Clothing ..	40 yards	yard	0·25	0·5
Firewood ..	10 maunds	maund	0·50	0·8
Kerosene ..	1 tin	tin	4	6
House-rent ..	..	house	12	15

(B. Com., Alld., 1949).
(M.A., Pb., Apr., 1951).

(32) Show with the help of the following data that the factor reversal test is satisfied by Fisher's Ideal formula for index number construction:—

Commodity.	Base year. price Rs.	Base year. Quantity	Current year. price Rs.	Current year. Quantity
A	6	50	10	56
B	2	100	2	120
C	4	60	6	60
D	10	30	12	24
E	8	40	12	36

(M. Com., Alld., 1946).
(M.A., Pb., Sept., 1951).
(B. Com., Delhi, 1953).

(33) The following table gives the index numbers for the several components of "Capital Index of Indian Industrial activity" with their respective weights for July 1940. Find out

(*a*) the simple, and (*b*) the weighted average index number of all the components and state why there is a difference between the two:—

	Index No. (Base 1935)	Weight
Indian Cotton Consumption ..	121·0	9
Jute Manufactures ..	127·6	6
Steel Ingots ..	143·8	5
Pig Iron ..	119·8	8
Paper ..	187·4	3
Coal ..	131·0	7
Rail & River-borne Trade ..	116·9	24
Cheque Clearances ..	85·2	20
Notes in circulation ..	116·2	6
Consumption of Electricity ..	135·8	7

(B. Com. Hons., Andhra, 1942).

(34) An enquiry into the budgets of the middle class families in a city in England gave the following information:—

Expenses on	food 35%	rent 15%	clothing 20%	fuel 10%	misc. 20%
Prices (1928)	£150	£30	£75	£25	£40
Prices (1929)	£145	£30	£65	£23	£45

What changes in cost of living figures of 1929 as compared with that of 1928 are seen?

(B. Com., Luck., 1944).

(35) Use the following data of industrial production in India to compare the annual fluctuations in Indian industrial activity by the chain base method:—

Index Nos. of Industrial Production in India.

Year	Index No.	Year	Index No.
1919—20	120	1926—27	149
1920—21	122	1927—28	156
1921—22	116	1928—29	137
1922—23	120	1929—30	162
1923—24	120	1930—31	149
1924—25	137	1931—32	160
1925—26	136	1932—33	160

(M. Com., Luck., 1943).

(36) From the fixed base index numbers in col. (2) below prepare chain base index numbers, and from chain base index numbers in col. (3) prepare fixed base indices:

Year	Fixed base index	Chain base index
(1)	(2)	(3)
1935	94	94
1936	98	104·3
1937	102	104·1
1938	95	93·1
1939	98	103·2
1940	100	102

(37) The following table gives *per Capita* income and cost of living index for India from 1939-40 to 1947-48:

Year	*Per Capita* Income (Rupees)	Cost of Living Index (Base: 1939-40)
1939—40	67	100
1940—41	70	105
1941—42	78	117
1942—43	112	160
1943—44	139	217
1944—45	139	216
1945—46	137	219
1946—47	143	242
1947—48	160	258

Deflate the money income with reference to the cost of living index, and represent the adjusted income diagrammatically for this period.

(M.A., Alld., 1950).

CHAPTER XII

DIAGRAMMATIC REPRESENTATION

An important function of the Science of Statistics is to present complex and unwieldy data in a manner such that they would be readily intelligible. Classification and tabulation constitute the first step towards the attainment of this objective; but even tables containing, as they do, a number of figures do not enable one to grasp the whole data at a glance. Computation of simple derivatives, statistical averages and index numbers constitutes further step in the direction of condensing the tabulated data. But still the condensed material is presented in numerical form. Numbers are not interesting to all. To many they are dull and confusing; and if their number is pretty large, it would be difficult to compare them and observe their differences. A long list of death rates and birth rates, to take an example, relating to a large number of towns in a country, or to different countries of the world, would tire one's eye and confuse his mind. It would not be easy for him to note the differences in death rates and birth rates of different towns, or countries as the case may be. Therefore, it is necessary to adopt a device which may present huge mass of quantitative data, or their condensed form, in a way that is at once comparable and appealing both to the eye and the intellect. For this purpose, the method of visual aids which comprises presenting statistical material in pictures, geometric figures and curves has been devised.

Usefulness of Diagrams.

Diagrams carry with them the merits of **attraction** and **effective impression.** One may not like to devote even a minute to the study of a page—a small page—containing a number of quantitative figures; and, even if he devotes time, numerical figures may go out of his mind soon after he has studied them. But the same person may not—in most cases, would not—like to take his eyes away from a picture relating even to the same topic to which the numerical data did. Nay, he might invite

others to have a look at the picture. And, if the picture has really attracted him, it need not be said that it would leave an effective impression on him. This is based on human psychology, and a successful advertiser or propagandist always exploits this psychology of the people to win his mark. A manufacturer of soap bars advertised for a considerable time that his bar, having the same price as that of his competitor, was much heavier than his competitor's but did not find any improvement in his sales. And, when he advertised in pictorial form—a balance containing his bar on one side and his competitor's on the other, the pan containing his touching the ground, while the other much above the ground, and the words "For the same price" beneath the picture—he found to his pleasant surprise that the demand for his bars increased so much that he had to extend his plant.

It follows from the above example that diagrams are not only attractive and impressive, but also have the merit of rendering the whole idea **readily intelligible.** A man, to take another example, who has never seen or handled more than a few hundreds of anything may not understand how large the biggest city of India, *viz.* Bombay, would be if he were told that its population in 1951 was 2,840,011 persons. But, if he were living in Kanpur and were told that Bombay was nearly four times of his own city in the size of population, he would, no doubt, *try* to understand what it means. The idea, however, shall be more easily and readily grasped if this fact were represented to him diagrammatically—*e.g.*, a circle may be divided into four equal parts, one of which may be shaded and named Kanpur, while the whole of the area would show the population of Bombay. It would be clear at a glance that Kanpur is only one-fourth of Bombay.

Another merit of diagrams is the ease with which they make **comparison** possible. Population of Kanpur with Bombay's, or the weight of one bar with that of the other, in the above examples, can be quite easily and readily compared.

Yet another characteristic feature of diagrams is that they **save much valuable time,** which would otherwise be lost in grasping the significance of numerical data.

Lastly, a chief merit of diagrams and graphs is that the entire data, which when expressed in numerical form may be

unwieldy and require a number of pages to write down, are made visible at a glance.

For these merits of theirs, diagrams are very useful in economic and social studies. A purely theoretical economist finds in them the basis for logical reasoning and easily explaining an economic law, such as the law of substitution or of diminishing utility. A practical economist may make his ideas impressive through diagrammatic representation. Knowing that the total imports of foodgrains into India amounted to only 2·1 million tons in 1950 as against the country's own production of 49·35 million tons of cereals and gram, he would do well to represent this fact diagrammatically rather than merely quoting the figures. When a social reformer is addressing an audience, mere reading out of figures would make the hearing dull, tedious and tiring. But, if he were to appear on the platform with picture, diagrams and graphs, his talk would be atonce interesting, lively and impressive. A businessman, or an administrator, has hardly any time to devote to the study of a huge mass of figures, however well arranged. But, if he is presented with graphs showing the rise and fall of a certain activity, or with pictures and diagrams, it will hardly take him more than a few minutes to grasp the significance of the whole. It is, thus, evident that diagrams, charts, pictures, graphs and similar other visual aids serve a more useful purpose than any other device.

But, diagrams can be as much misused as they are useful. In advertisements and political propaganda they have often been found to be misleading. The true statistician has to guard himself against mis-representation. Hence some general directions for drawing diagrams.

Directions for Drawing Diagrams.

It should be remembered that diagrams do not add anything to the meaning of statistics. They afford only a method of presenting the figures They are not intended to connote more than the data they represent. They prove nothing. Drawn and studied intelligently, they bring to light the features of statistical groups and series; they show the various components of a group in relation to each other and to the group as a whole; they show the unity that underlies the scattered

figures. They are, therefore, only a means to an end, the end being to make comparisons. Consequently, if there is only one isolated numerical quantity, there is no sense in presenting it diagrammatically. Similarly, if there are many figures, in no way related to one another and, therefore, having no common characteristic, they are incomparable and, therefore, need not be diagrammatically presented. For example, if we know that the monthly expenditure of a certain student is Rs. 50, his age is 22 years, the length of his nose is 1.325 inches and he has 20 books, we cannot represent Rs. 50, 22 years, 1.325 inches and 20 books by any kind of diagram, since the four numbers are incomparable. On the other hand, if we know that one student is 60 inches long and another only 48 inches, we can compare the two measurements and, therefore, represent them diagrammatically. It is, then, established that *the method of diagrammatic representation can be made use of when there are at least two numbers which are similar in nature and character at least in one important respect and which also vary independently of each other.*

Another point that should be kept in view is that *diagrams are not the substitutes for the real magnitude of the quantity they represent.* The size of a diagram changes with the change in the scale to which it is drawn. The same quantity drawn to two different scales will yield diagrams of different sizes.

It is evident from the above that in the technique relating to drawing of diagrams the selection of the proper scale occupies an important place. No rigid rules can be laid down for the selection of a proper scale, but a general direction that can be given for the purpose is that that scale is the most suitable the diagram drawn to which would be neither so big as not to be visible at a glance nor so small as to look clumsy and indistinct and cover only a very small part of the space available. The scale should be so chosen that the size of the resulting diagram would show the significant features of the numerical quantities for which it stands. All principal details must be clear. The vertical scale should be marked at equal unit spaces and the measurement of each unit space put down. Generally, the vertical scale should be shown on the left-hand side of the diagram. The horizontal scale should be given at the bottom

of the diagram. On each side, the vertical and the horizontal, the variable represented should be indicated. For instance, the vertical line might show the amount in rupees and the horizontal the different countries. The diagram should be neatly drawn with the help of drawing instruments. It should be given a suitable heading. The data represented diagrammatically should be given on a page adjacent to the one on which the digram is drawn. If these data are indicated in the diagram itself, care should be taken to see that the quantities are so placed in the diagram that they do not distort the visual impression conveyed by the diagram. To make distinctions clear, various kinds of dotting, lining, crossing, cross-hatching, or colouring should be used.

The drawing of diagrams, as a matter of fact, is not so difficult as the selection of suitable types of diagrammatic forms to depict a concise picture of the statistical data in hand. In selecting the most suitable diagram from among the varied forms of diagrams, the criterion should be that the diagrams selected should lead most *quickly* and with the greatest *accuracy* to the real meaning of the quantitative data. The test of a successful selection lies in the speed with which the quantities can be accurately studied with the help of the diagram selected.

Difference Forms of Diagram.

Diagrammatic representation can be made in any one of the following ways:

(1) One dimensional diagram, *e.g.*, lines or bars drawn to a common scale.

(2) Two dimensional diagrams, *e.g.*, squares, and rectangles whose areas are made proportional to the given figures.

(3) Circular and angular diagrams, *e.g.*, circles whose areas are made proportional to given magnitudes, and which may be divided into sectors whose common unit is the degree.

(4) Three dimensional diagrams, *e.g.*, cubes, cylinders, blocks whose volumes are made proportional to the given figures.

(5) Pictograms, *e.g.*, statistical maps and pictures.

Technically there is no objection to using squares, rectangles, cubes, circles and pictures; but in practice, lines, bars and angular diagrams are the easiest to draw. They can also be made sufficiently accurate. Therefore, so far as possible they should be preferred.

The terms diagram, chart, and graph are very often used without distinction, the same figure being given any of these names. We shall use the term 'diagram' for the various forms pointed out above, and the term 'graph' for curves only, which would be dealt with in later chapters.

One Dimensional Diagrams—Simple Bar.

A bar is merely a thick line whose width, though shown in the diagram, is not taken into consideration. It is shown merely to make the diagram look attractive. Therefore, those diagrams in which only one dimension is considered are called one dimensional diagrams. Now, a bar may be shown as a simple bar or it may be divided into parts. That one dimensional diagram in which the bars are not sub-divided, that is, where the bars represent the whole of a magnitude and not components of it, is called a **simple bar diagram.** We consider it first.

To draw a bar diagram the height of the biggest bar should be adjusted to the size of the diagram. Some margin should be left all round the diagram to write down the title and the designation of units and scale. The width of the bars should be neither too big nor too small. If the number of items is very large and the space very limited, thick bars may be replaced by thin lines as done in Figure 22. The bars should be drawn to a common horizontal or vertical base line; generally the horizontal base is used on which bars are made to stand vertically. Horizontal base is used for the simple reason that comparison of one bar with another can better be made in terms

of height. In a single study all the bars must be of the same uniform width, separated by equal intervening spaces. Bars may be coloured, lined or dotted, but the colour used or lining or dotting done should be the same in all the bars in a single study. If bars are made to touch each other, that is, no intervening blank spaces are left, the diagram would look a continuous and blurred one with its top disfigured. Bar diagrams are not suitable for presenting continuous series such as that spread over a long period of time. For this purpose graphical methods of presentation are used, which will be studied in later chapters. Bar diagrams art suitable for representing spatial series.

Table 34 gives the yield of certain food crops in India (prior to partition) for 1936-37,[1] which are diagrammatically represented in Figure 3.

Table 34. *Yield of certain food crops in British India (including minor States)*—1936-37.

Crop	Wheat	Sugar-cane	Jowar	Gram	Barley	Maize
Yield (1,000 Tons)	8,513	6,289	5,401	3,817	2,311	1,836

The highest yield to be represented is 8,513,000 tons. A suitable scale has been selected to represent this yield properly, and the yield of other crops has been reduced to this scale. The lower ends of all the bars have been placed on the common horizontal base, so that comparison can be made between yields of different crops by comparing their heights. To make comparison easy, the bars have been arranged in descending order; they could have been arranged in ascending order as well. The scale is put down on the left-hand side, a little away from the biggest bar. Names of the crops are given below each bar. A suitable heading is given at the top. The bars are separated by equal blank inter-spaces.

[1] Thomas and Sastry, *Indian Agricultural Statistics*, page 119.

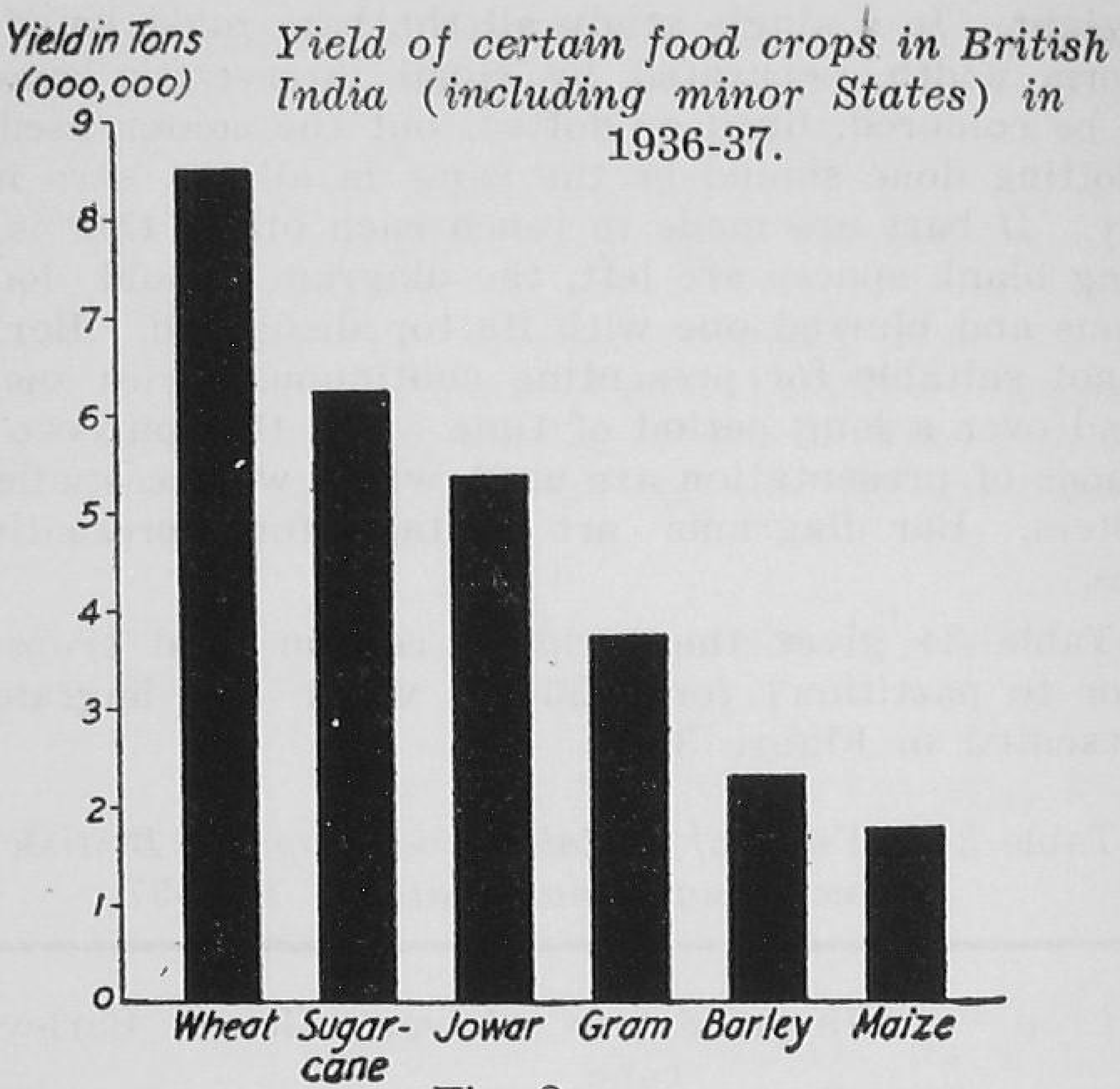


Fig. 3

In interpreting Figure 3, it should not be said that in 1936-37 the yield of wheat in India was the highest among *all* the crops, for there might be some crop other than the crops shown in the diagram whose yield may have been higher. Actually, the yield of rice in the same year was 27,143,000 tons. It can, of course, be said without any mistake that among the six crops represented in the diagram the yield of wheat was the highest and that of maize the lowest in 1936-37.

In Figure 22, heights of 55 boys have been shown by vertical lines, the method of construction being nearly the same as that of Figure 3.

If instead of yield of food crops in India, we were given the imports, death rates, or income *per capita* of different countries of the world the above method of simple bars would have been followed to represent them diagrammatically. Again, if we were given imports and exports of different countries and it was desired to compare imports of different countries, exports of

different countries and imports and exports of the same country, we could have extended the principle of simple bar diagram for this purpose as well. Choosing the horizontal or the vertical line as the base we could have placed two bars, one representing imports and the other exports of the same country, and separated this pair by an intervening blank space from another similar pair for another country, and so on, colouring the bars showing imports with one colour and those showing exports with another colour. This method is somewhat complex for simple comparisons, and can be replaced by that of sub-divided single bars.

One Dimensional Diagrams—Sub-divided Bar.

If a given magnitude can be broken up into the parts of which it is composed, or, if there are independent quantities constituting the sub-divisions of a total, bars sub-divided in the ratio of the different components may be used to show the *relationship of the parts to the whole.* For instance, if death rates and birth rates of different countries are given, bars may first be drawn to represent the births, and, from these bars, portions from the bottom of the bars may be cut out in proportion to the death rates and coloured black to distinguish from the remaining white portion which would show the survival rate. Again, if the population of a number of countries is given, bars representing the population of different countries may be further sub-divided into two parts in proportion to males and females, one portion being shaded black to distinguish from the other. Technically, in these two examples we would say that the bar showing death rate has been super-imposed on that showing birth rate, or bars showing males and females have been super-imposed on the bar showing the total population.

Table 35 gives the value of exports and imports of India (prior to Partition) in total merchandise for three years. These exports and imports may be added up, and bars proportionate to the totals may be drawn, the three bars being of unequal height because of the inequality of the total sea-borne trade. These bars may then be sub-divided into two portions, the lower one in each one of the three bars showing the exports and painted black and the upper portion, remaining white, showing the imports. Thus three comparisons will be possible at a glance

—*viz.*, those relating to exports, imports and total trade in different years.

Another method may be to draw three bars of equal length to show the total foreign trade which in each case may be put down at hundred; these bars should then be subdivided according to the percentage which exports and imports in each year bear to the total foreign trade of that year.

Let us suppose that the three bars are 5 inches in length each. The total foreign trade for 1923-24 is about Rs. 600 crores, and the exports and imports are respectively Rs. 363 crores, and Rs. 237 crores, so that they are respectively 60 per cent. and 40 per cent. of the total foreign trade. The bar of 5 inches for 1923-24 would, therefore, be divided into the proportion of 3: 2 to represent 60 per cent. and 40 per cent. respectively. This single bar now represents percentage values of the imports and exports to total foreign trade of India separately. The difference between this method and the former method should be carefully noted. The former method makes possible the comparison of actual values of imports with exports and of imports or exports with the total foreign trade; while the latter method makes possible the same comparisons in percentage values.

If, however, the aim is to show the balance of trade the method of sub-divided bar diagram can be applied to subdividing either the bar for exports or that for imports, whichever is greater, into two portions, one representing the imports or exports whichever is less and the other showing the balance of trade, positive or negative, as the case may be.

Table 35. *Value of Sea-borne Trade of India in Total Merchandise (including Govt. Stores).*

Year	Imports	Exports	Balance of Trade
	Crores of Rs.	Crores of Rs.	Crores of Rs.
1921-2	282.59	248.65	−33.94
1922-3	246.19	316.07	+69.88
1923-4	237.18	363.37	+126.19

To represent the figures of exports, imports and balance

of trade given in table 35 a common horizontal base line is chosen in Figure 4. For 1922-23 and 1923-24 two bars with their heights in proportion to the respective amount of exports are drawn; then, from the bottom of the two bars heights in proportion to imports are cut off and painted black. The heights of the full bars represent the values of exports, coloured portions represent the values of imports, and the remaining white portions show the favourable balance of trade indicated by plus sign in column 4, table 35. For the year 1921-22 the imports are greater than the exports. Therefore, first a bar representing the value of imports for the year is drawn on the same vertical scale. From the bottom of this bar a height equalling exports is cut off and left blank, the small portion at the top being coloured black. The height of the full bar indicates total imports, white portion in it represents exports and the coloured portion at the top shows unfavourable balance of trade. Were, however, the exports in a particular year exactly equal to the imports, the bar would be painted black, without the necessity

Value of sea-borne trade of India in total merchandise (including government-stores)

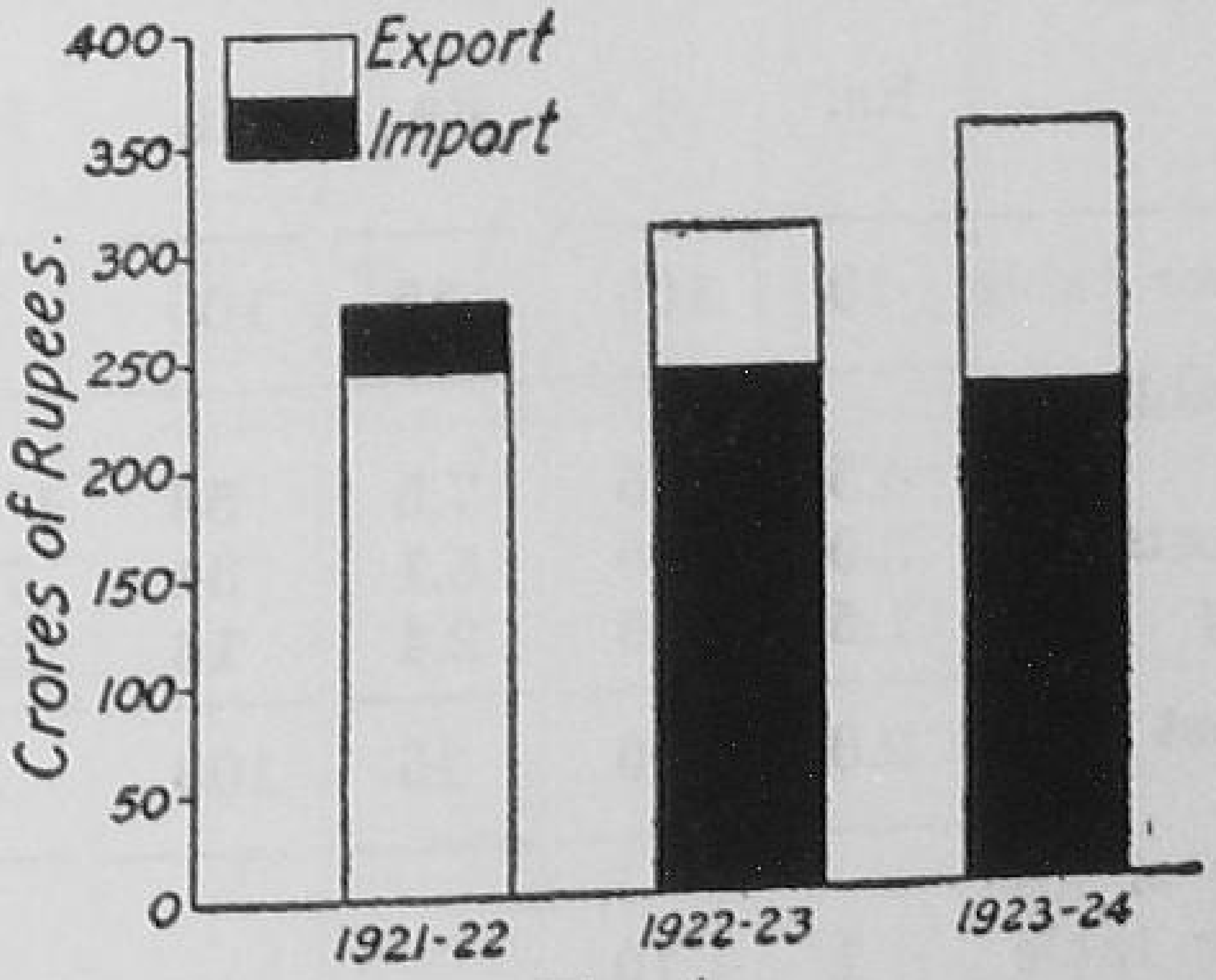


Fig. 4

Note.—The shaded portion at the top of the bar for 1921-22 indicates unfavourable balance of trade, while the white portions at the top of the bars for 1922-23 and 1923-24 indicate favourable balance of trade.

of being sub-divided. The height of the bar would indicate total imports as well as total exports, meaning that the balance of trade was nil. Figure 4 makes possible the comparison of actual values. Comparison of percentage values can also be made by drawing the bars on percentage basis in the manner indicated above.

A single bar may be sub-divided into more than two sub-divisions as well. Table 36 gives the proceeds, cost, and profit and loss per table during three years. In the year 1938 there is a gain of Re. 1; in 1939 there is neither gain nor loss; while, in 1940 there is a loss of Re. 1. It is desired to represent the given data by sub-divided bar diagram drawn on percentage basis.

Table 36. *Proceeds, Cost, Profits or Loss, per table during 1938, 1939 and 1940.*

Particulars	1938		1939		1940	
	Rs.	%	Rs.	%	Rs.	%
Proceeds per table	10	100	15	100	20	100
Cost per table—						
Wages	4.5	45	7.5	50	10.5	52.5
Other Costs	3.0	30	5.1	34	7.0	35.0
Polishing	1.5	15	2.4	16	3.5	17.5
Total Cost	9.0	90	15	100	21	105.0
Profit (+) or loss (−) per table	+1	+10	..	..	−1	−5

In table 36 the percentages of wages, other costs and polishing to proceeds per table are shown for the different years. In 1938 the profits is 10% of the proceeds and in 1940

the loss is only 5% of the proceeds, although the actual profit and loss in both the cases are the same, *viz.*, one rupee. Subdivided bar diagram drawn on the percentage basis would, therefore, be better in so far as comparisons of relative values are concerned; but it would not be suitable if actual values are to be compared. Figure 5 shows the diagram on percentage basis.

Percentages of cost of, and profit and loss on, a table in 1938, *and* 1940.

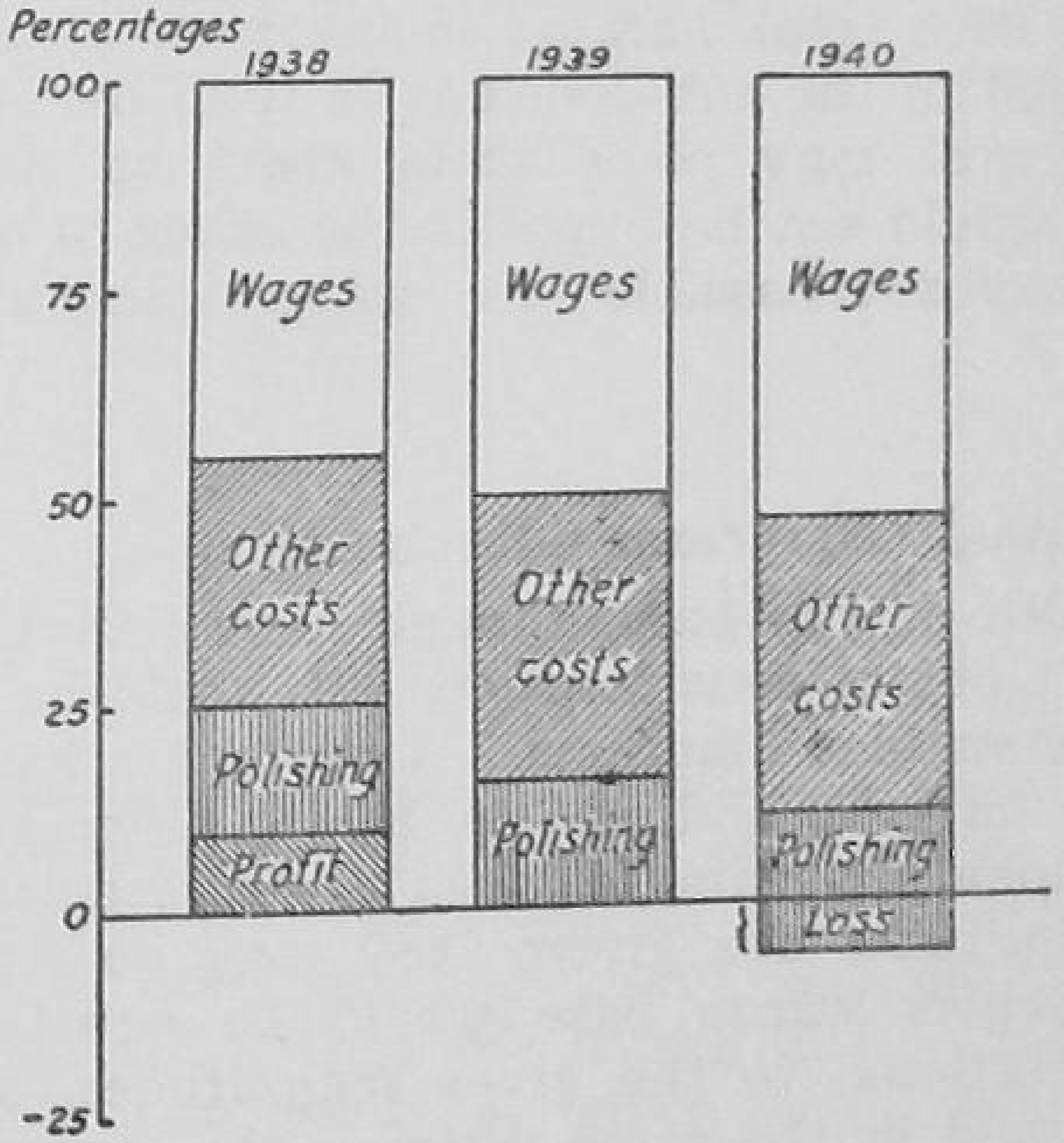


Fig. 5.

The construction of the diagram to represent the data in table 36 should be carefully studied. First, three bars to represent the proceeds per table are drawn and made equal to 100. The proportions of wages, other costs and polishing are then cut down from the bars such that the same order

for each of them is maintained in the bars. The surplus in the bar for 1938, just above the horizontal base line, indicates profit; there is no surplus in the bar for 1939. In the bar for 1940 the deficit of 5% has been made good by extending the bar below the horizontal line in minus direction, the portion below the horizontal line showing the loss. Comparison in terms of percentages can be easily made from the figure. Bars could also be similarly drawn to compare actual values.

Although bars may be used for showing the sub-divisions of a large number of totals, it is not advisable to adopt the bar method for comparison if there are more than three or four components of each total, because in that case even if the same order is followed in the sub-divisions in each bar, the disparity among the figures may place them apart so that one type of component would not be opposite the other in all cases, and therefore it may not possible to make comparisons at a glance.

Two Dimensional Diagrams—Rectangles.

The breadth of the bars, though shown in the diagrams, was so far left out of consideration. It would be utilised now in drawing rectangular diagrams. A rectangle has two dimensions. Hence, its area, and not the height alone as in the case of one dimensional diagram, is taken to represent a magnitude. If several magnitudes are given, they may be represented by separate rectangles whose base would be equal but heights would be proportional to the given magnitudes, so that their areas would stand in the same ratio as the given magnitudes. Rectangles are suitable for use in cases where two or more quantities are to be compared and each quantity is sub-divided into several components.

Table 37 gives the monthly incomes of two families and their expenses on different items. The data would be properly represented by rectangular diagram. The incomes of the families are Rs. 80 and Rs. 40. Therefore, in Figure 6, on the same horizontal line two rectangles are erected such that the widths

of the two are in the ratio of 80: 40, or 2: 1, and the heights are equal. Thus the areas of the two rectangles are in the same proportion as the incomes. Each of the rectangles is, then, sub-divided into components according to the percentage

Table 37. *Family-budgets of two families.*

Items of expenditure	Family A, Income Rs. 80			Family B, Income Rs. 49		
	Actual expenses	Percentage	Cumulative percentage	Actual expenses	Percentage	Cumulative percentage
	Rs.			Rs.		
1. Food	32	40	40	20	50	50
2. Clothing	20	25	65	8	20	70
3. Shelter	8	10	75	4	10	80
4. Fuel and light	4	5	70	2	5	85
5. Miscellaneous	16	20	100	6	15	100
	80	100	..	40	100	..

of income spent on different items of expenditure. Percentages are shown in table 37. In order that the marking out of sub-divisions may be easy and convenient, cumulative percentages have been calculated. They are shown in the table in columns 4 and 7. The vertical scale is given on the left, and with its help cumulative percentages have been marked in the rectangles. Thus, for family A the first mark is put at 40, the second at 65, the third at 75, the fourth at 80, so that the remaining 20 is the percentage expenditure on miscellaneous items. Cumulation of percentages removes the chances of error in marking the sub-divisions.

The areas of the components of the rectangles are proportional to the actual expenses on various items, and the areas of the rectangles are in proportion to the income. Therefore, comparison of percentages of income spent over different items in the same family and in the two families is rendered easy. A

Percentage of income spent by two families on different items of expenditure.

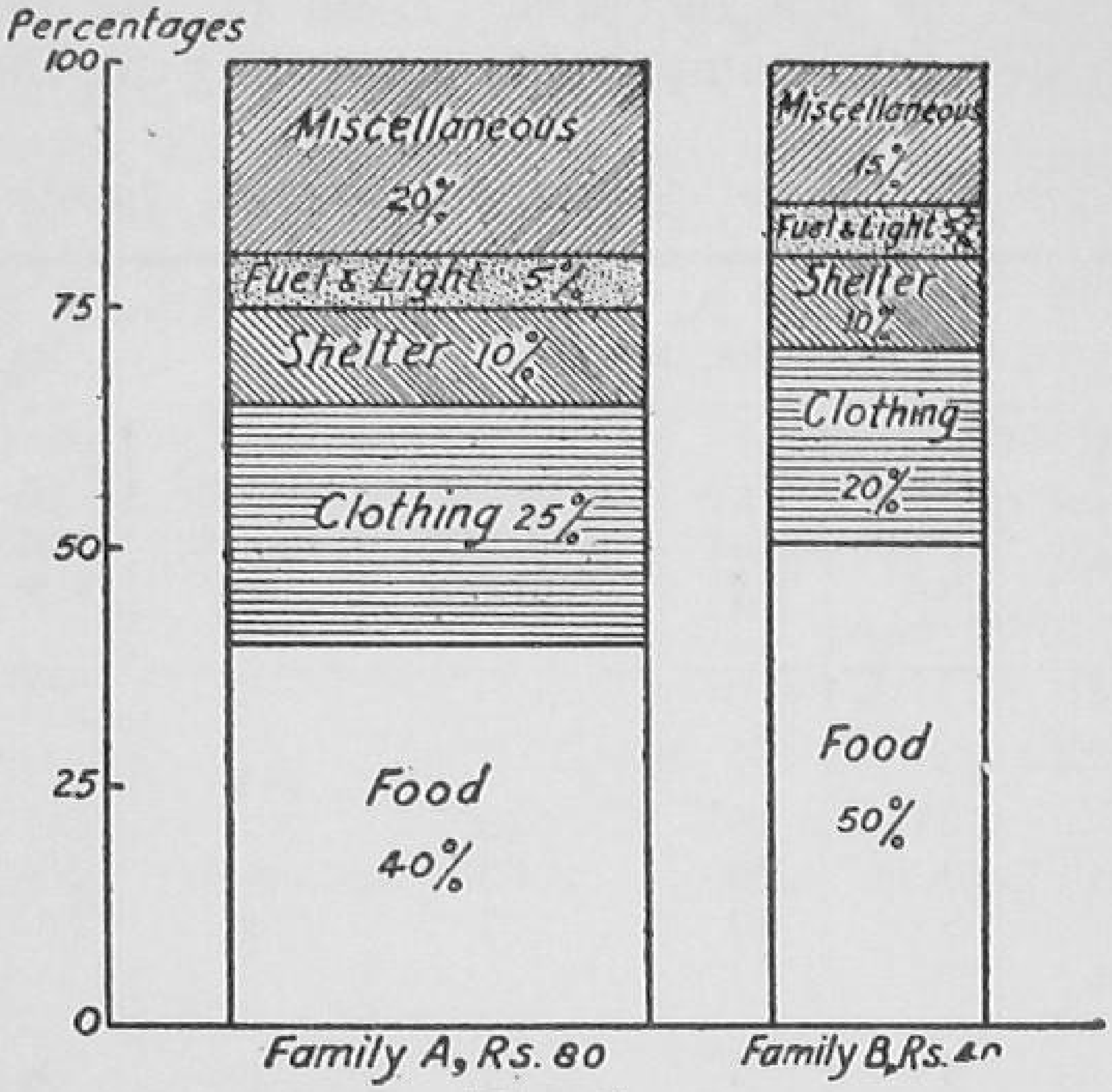


Fig. 6.

glance at the heights of the rectangles relating to percentage expenditure on shelter and fuel and light would show that these heights are identical in the two. It implies that the percentage of income spent on these items in the two families is equal but the actual expenditure is different. The actual expenses stand in the same ratio as the areas of the components. Similarly, there might be a case in which the heights of the components may be different but their areas equal, implying that the actual expenses are identical but percentage expenditure is not, the percentage expenditure being in the same ratio as the heights of the components.

Rectangles are not used to represent only the family budgets. They are two dimensional diagrams and can, therefore, be also employed to show three different factors. Consider table 38. It gives (1) price of a commodity, (2) its quantity sold and (3) different expenses of production and net profit. These facts have been diagrammatically represented in Figure 7, a close study of which would reveal that the heights of the

two rectangles are in proportion to the quantities sold, and the widths are in proportion to cost per unit, so that the areas of

Table 38. *Details of price, cost and quantity sold of two commodities.*

Price of a Commodity		I. Rs. 2 per unit	II. Rs. 3 per unit
Quantity Sold	..	40	20
Value of raw-materials	..	Rs. 26	Rs. 24
Other expenses on production	..	Rs. 32	Rs. 21
Profits	..	Rs. 22	Rs. 15

the rectangles are in proportion to the total proceeds in each case. These areas are sub-divided into their components, *viz.*, expenses on raw materials and production, and profit.

Details of Cost of two commodities.

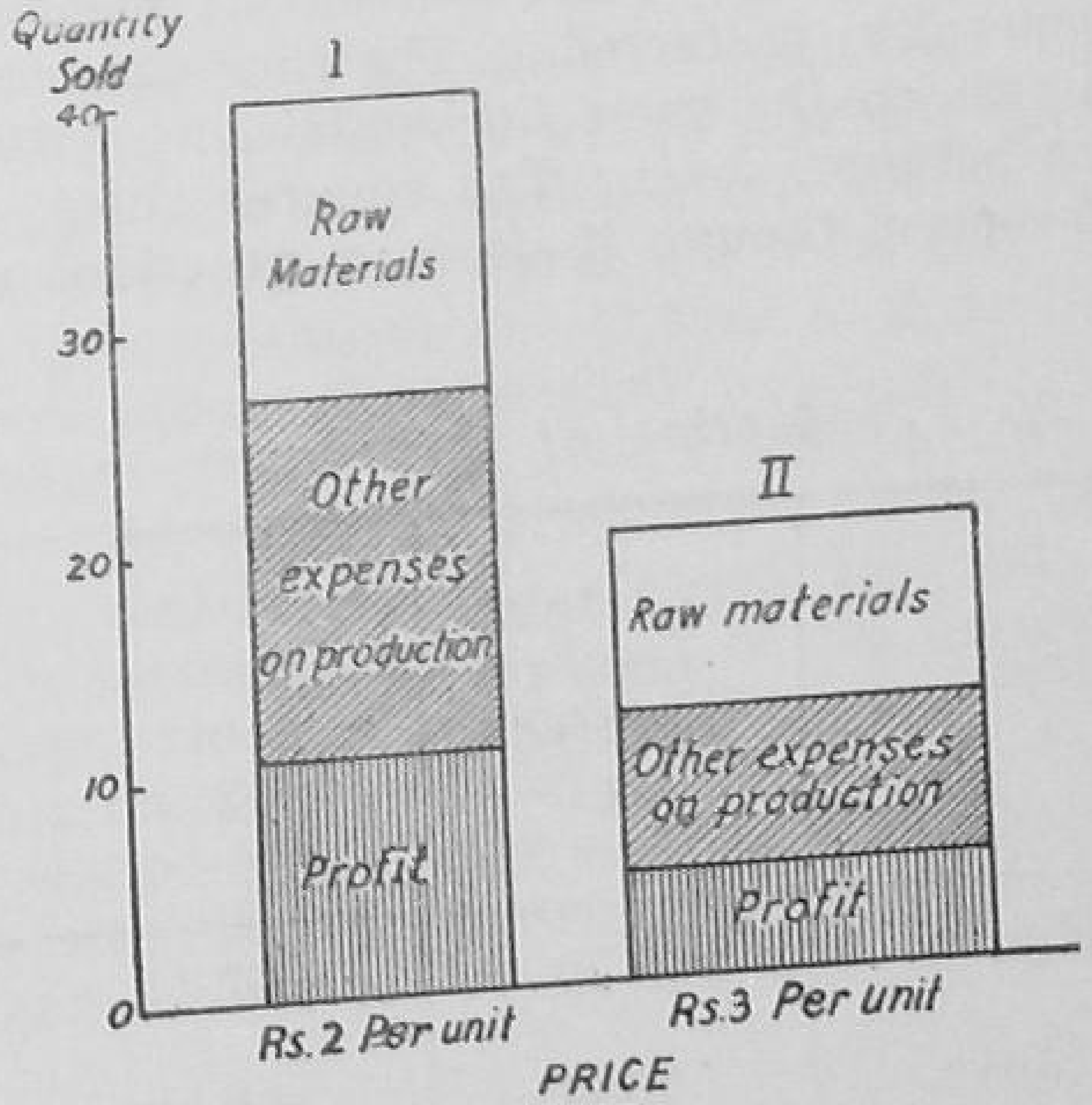


Fig. 7.

It should be noted that rectangles are the only two dimensional diagrams which are capable of showing three different factors. Squares and circles which are also surface diagrams do not enjoy this property. Hence, the place of rectangles in diagrammatic representation is very important.

Two Dimensional Diagrams—Squares.

When it is desired to compare quantities which bear the ratio of 1:100, bar diagrams fail to answer the purpose, since however small the scale selected be, the height of one bar will have to be made 100 times as great as that of the other, with the result that one bar will be too tall and the other too small. In such cases bars are replaced by squares.

The side of a square varies as the square-root of its area. If, therefore, two figures 100 and 10,000 are to be represented by squares, their sides would be taken in the proportion of $\sqrt{100}$: $\sqrt{10,000}$ or 1:10 and not in that of 100: 10,000.

Table 39 gives the production of cane-sugar in four countries in 1938-39, arranged in descending order. Column 2 of the same table gives the approximate square-roots of figures given in column 1. Column 3 contains numbers which are obtained by dividing the square-roots by 50. These numbers give the sides of the squares. They can be taken to be so many inches, cms. or any other convenient measure of length. Inches are generally preferred. The four squares in Figure 8 are made to rest on the same horizontal line with equal intervening spaces between them. The squares show the proportionate differences, though it needs a practised eye to detect them.

Table 39. *Production of Cane-sugar in* 1938-39.

Countries	Quintals oooo's omitted 1	Square-roots 2	Sides of the squares in inches 3
1. India (undivided).	2750	52.44	1.05
2. Neth. Ind: Java ..	1550	39.37	0.79
3. Hawaii ..	835	28.90	0.58
4. Columbia ..	51	7.14	0.14

It will be seen that scale is given above the squares in Figures 8. Let us see how this scale is calculated. The side of the first square, relating to India, is 1.05 inches. The area of the square is, therefore, 1.1025 square inches. It represents 2,75,00,000 quintals. Therefore, 1 sq. inch represents 249,40,000 quintals nearly. This is the scale to which the squares have been drawn in Figure 8.

Cane-sugar production in certain countries, 1938-39.

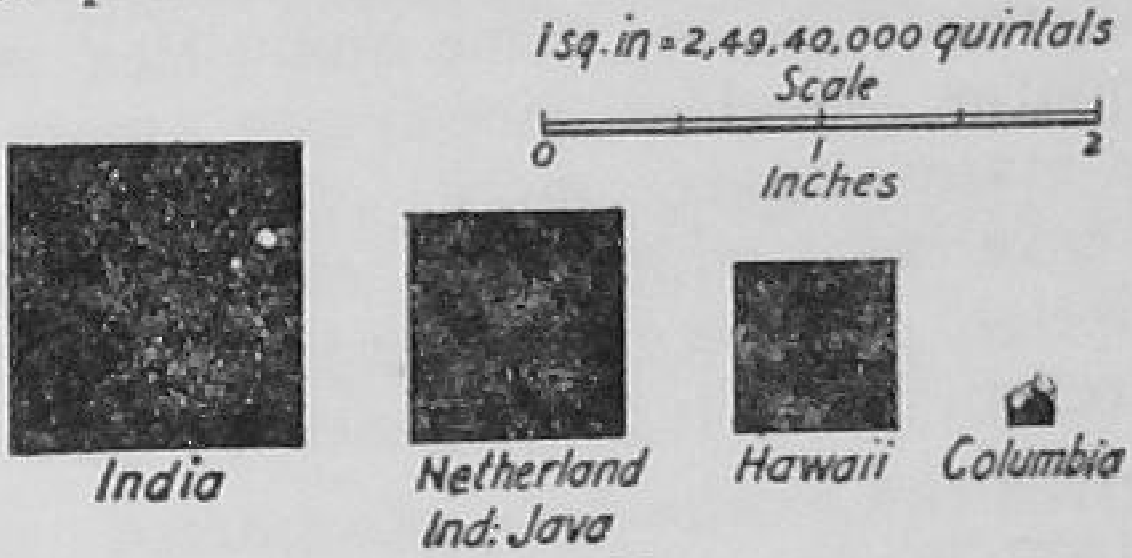


Fig. 8.

In Figure 8 squares are shown separately one from the other. Separate squares do not afford a proper view of proportions at a glance and also require much space. If a total is capable of being divided into parts, the total may be shown by a square and its components sub-divided in it as they are sub-divided in rectangles. This method would not require much space, and would, at the same time, render comparison of proportions easy. For instance, given the area of the world and of the several continents, the area of the world may be represented by a square and the areas of the several continents by its rectangular sub-divisions. The square may be sub-divided horizontally or vertically.

Squares require much time and labour to be drawn accurately. They may, therefore, be replaced by circular diagrams. Circles take less time to be drawn, and can be drawn sufficiently accurately.

Circular Diagrams—Circles.

The area of a circle varies as the square of its radius. If the radius of a circle is twice that of another, its area would be four times the area of the other. Similar is the case with

squares. If the side of a square is twice that of another, its area would be four times the area of the other. It follows that if the radii of several circles are in the same proportion as the sides of the squares, the areas of the circles would also be in the same proportion as the areas of the squares. Therefore, the same quantities may be used either as the radii of several circles or as the sides of several squares, without resulting in any difference in their comparative study. Thus the squares shown in Figure 8 may be replaced by an equal number of circles whose radii would be the same quantities as represent the sides of the squares. Besides, circles are more attractive to look at and, in many cases, more effective to compare than squares. The ease with which they can be drawn is another advantage of circles. Therefore, when there is a choice between squares and circles, the latter are generally preferred.

Table 40 gives the production of ginned cotton in five countries, for 1937-38, in column (2). The square-roots of

Table 40. *Production of Ginned Cotton,* 1937-38.

Country (1)	Production in Quintals oooo's omitted (2)	Square roots (3)	Radii in inches (4)
India (undivided)	1048	32·3	1·53
China ..	700	26·4	1·27
Brazil ..	460	21·5	1·02
Peru ..	81·5	9·0	0·43
Argentine ..	51·4	7·2	0·34

figures in column (2) are given in column (3). These square-roots may be taken as the radii of the circles, but since they are too big to be shown in inches or cms., they are divided by 21. The numbers thus obtained are shown in column (4).

They stand for radii in inches, and have been used in drawing the circles in Figure 9. The circles are arranged in descending order. The proportions of production of ginned cotton in diffe rent countries are comparable at a glance. In interpreting the circles in Figure 9 we should not say that production of ginned cotton is the least in Argentine in the world, since there may be several other countries, not shown in the diagram, which may have still less production. All that can be said is that among the five given countries production in Argentine is the least.

Production of Ginned Cotton in certain countries.

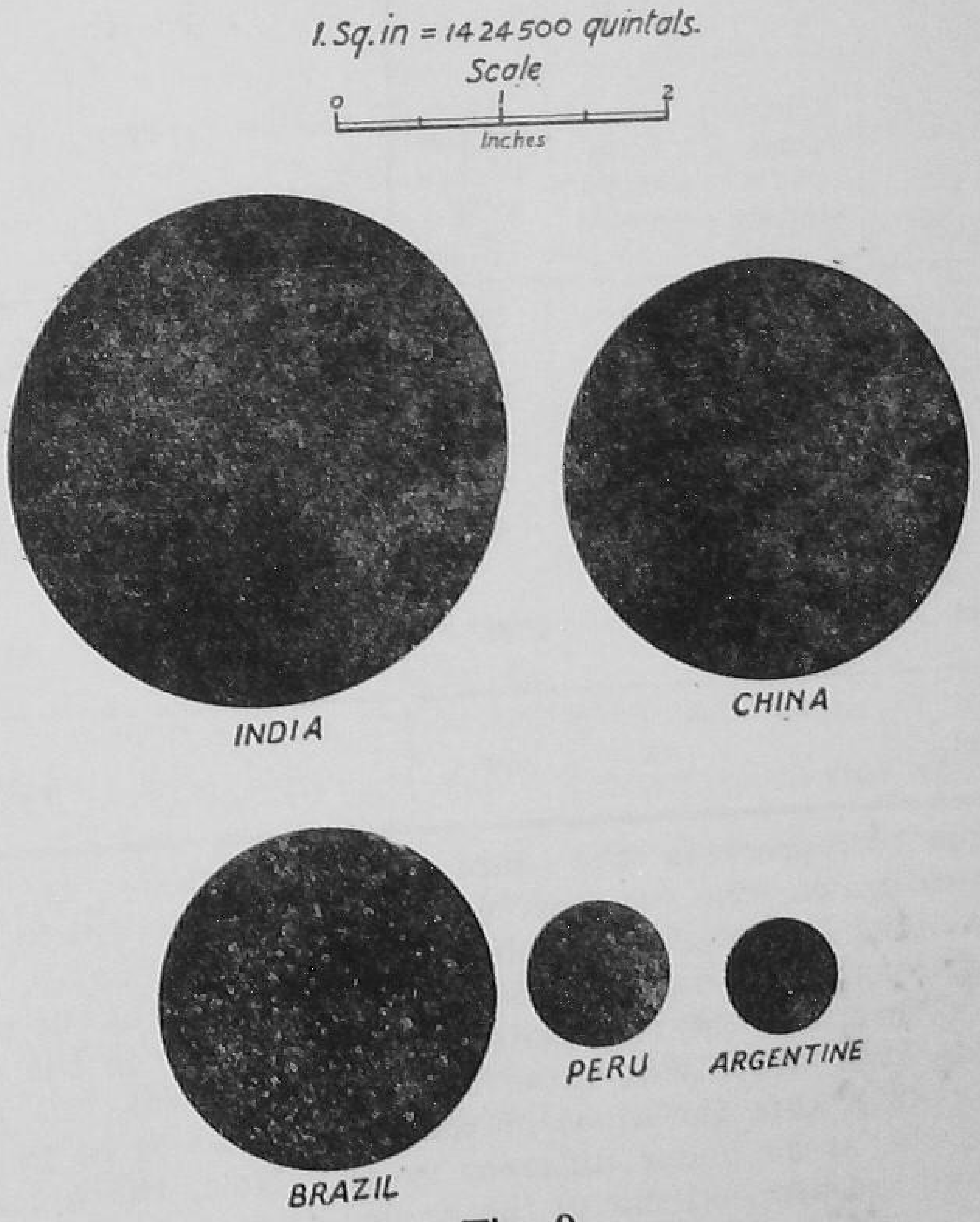


Fig. 9.

Angular Diagrams—Sectors.

Just as a rectangle or a square can be sub-divided into its components, a circle can also be divided into sectors to represent the parts of a total. The method of circles and sectors is to be preferred to that of sub-dividing a square, because of the suggestive and attractive character of sectors and circles, and also because of the ease with which they can be drawn.

Table 41 gives the area under food-crops cultivated in 1939-40 in U.P. and Punjab (undivided).

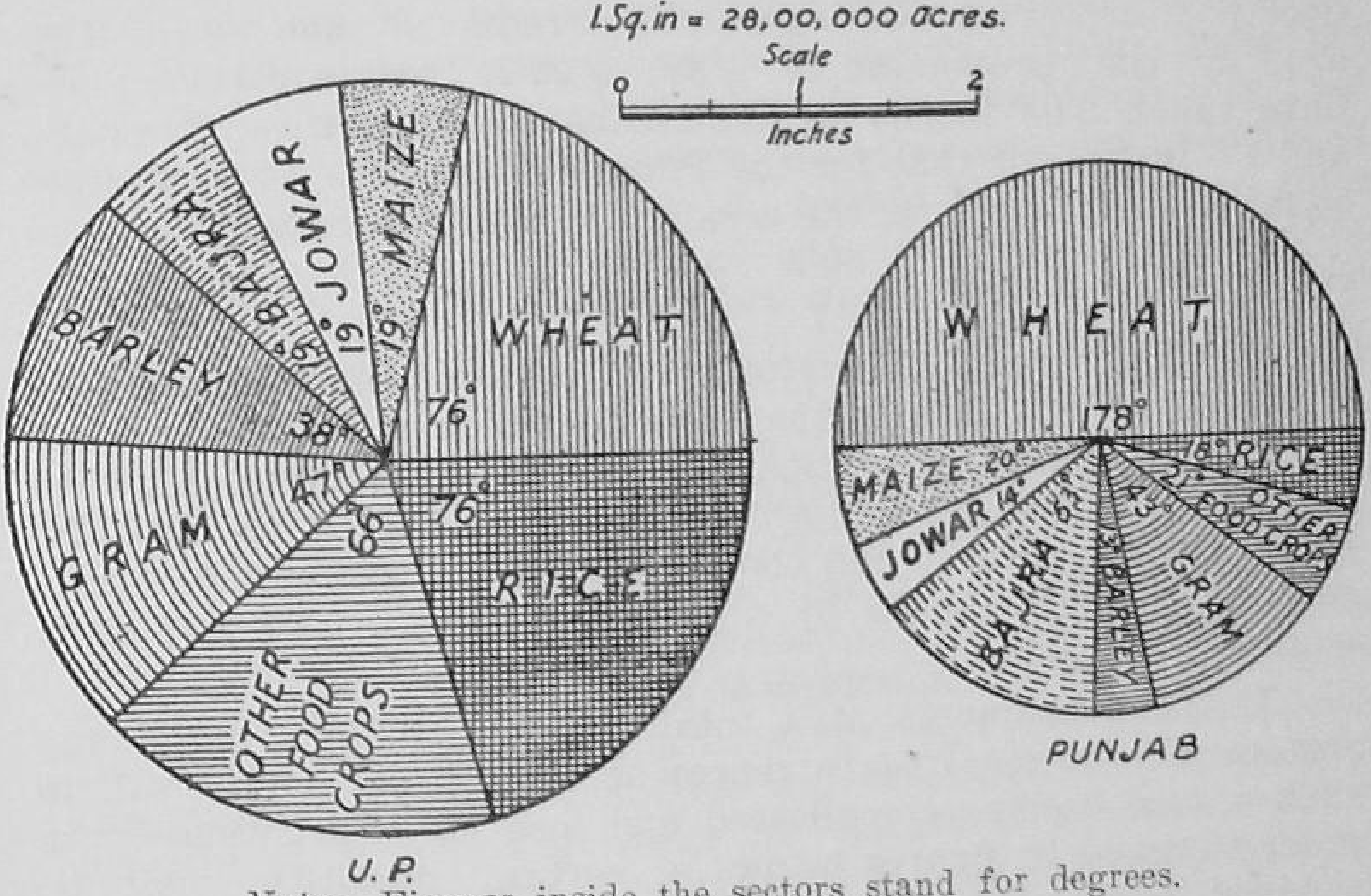
Table 41. *Area under food-crops in U.P. and undivided Punjab in 1939-40.*

Food crop	U.P.			Punjab		
	Area in acres 000's omitted	Area in acres 000,000's omitted	Angles of the sectors	Area in acres 000's omitted	Area in acres 000,000's omitted	Angles of the sectors
Rice ..	7764	8	76°	977	1	80°
Wheat ..	8109	8	76°	9566	10	178°
Barley ..	3823	4	38°	730	0.7	13°
Jowar ..	2307	2	19°	778	0.8	14°
Bajra ..	2388	2	19°	3061	3	53°
Maize ..	2107	2	19°	1143	1.1	20°
Gram ..	5399	5	47°	2413	2.4	43°
Other food grains & pulses ..	6819	7	66°	1231	1.2	21°
Total ..	..	38	360°	..	20.2	360°

Figue 10 represents the data given in table 41. Two circles are drawn, one for area under total food-crops in U.P. and the other for area under the same crops in Punjab, such that the areas of the two circles are in proportion to the areas under food-crops in the two States. Their radii are in the ratio of $\sqrt{38}:\sqrt{20{\cdot}2}$. To divide these circles into sectors, the principle to note is that the areas of the sectors should be proportional to the areas under different crops. Now, the areas of the sectors are proportional to the angles at the centre. There-

fore, 360°, the total number of degrees contained in a circle, are to be divided into proportional parts to get sectors of the required areas. 38 and 20.2, in the example under consideration, are put down as equal to 360° and, by the simple rule of three, the number of degrees contained by the angle of the sector representing the area under any crop is calculated.

Area under different food-crops in U.P. and the Punjab.



Note.—Figures inside the sectors stand for degrees.

Fig. 10.

Thus, the angle of the sector for area under rice in U.P. is equal to $\frac{8}{38} \times 360 = 76°$ nearly. Similarly, angles of other sectors standing for other crops in the two States are found out. Having determined the angles of different sectors, we show them at the centre of circle. The process would be to draw any radius in a circle, and taking it to be the starting point, to mark off the requisite number of degrees on the circumference with the help of the protractor, and then to draw radii from these marks to divide the circle into requisite number of sectors.

It should be noted that we have proceeded with approximated figures in calculating the angles of the sectors and not with actual figures. The difference between angles found by using the correct and the approximate figures would be insignificant and immaterial for all practical purposes. If greater accuracy is desired, the direct method where actual figures are used should be adopted.

Scale is given at the top in Figure 10. Its calculation does not involve any difficulty. The radii of the two circles are in the proportion of $\sqrt{38} : \sqrt{20.2}$, that is $6.1 : 4.5$. We have taken 2·03 inches as the radius for the circle for U.P., and 1·5 inches for that for the Punjab. The area of a circle is equal to πr^2; therefore, the area of the circle for the Punjab is equal to $\frac{22}{7} \times (1.5)^2 = \frac{50.5}{7}$ square inches. It represents 2,02,00,000 acres. Therefore, 1 square inch represents 28,00,000 acres. This is the scale to which circles have been drawn. In Figure 10, comparison of the areas under all food crops in U.P. and the Punjab can be made at a glance by looking at the two circles, and comparison of the areas under different crops can be made by a glance at the sectors and their angles.

If the components of a total are too many, too many sectors will be required to represent them. The circle will in such a case become complicated and lose its effectiveness. In order to avoid it, figures below a certain quantity may be grouped together so that the number of sectors may be reduced to manageable limit.

In Figure 10, it will be seen that in the circle representing the area under food-crops in U.P., the sectors are arranged in order of magnitude. This arrangement has the advantage that even if minute differences in areas for different crops are ignored in the calculation of the angles of the sectors, they will be clearly indicated by the order in which the sectors occur in the circle. For instance, the areas under rice and wheat in U.P. are, respectively, 77,64,000 and 81,09,000 acres. The difference is very little and has to be ignored in the process of calculating the degree of sectors; but this difference is accounted for in the circle by placing

wheat before rice. Similarly, the differences between the figures for *jowar, bajra* and maize are so slight that they give equal degrees for their sectors, but by placing *bajra* before *jowar*, and *jowar* before maize the fact, that although their degrees are the same their magnitudes are in the order in which they have been arranged in the diagram, is made clear. Again, it will be noted that the arrangement adopted in the second circle, namely, that for the Punjab, is just the same as that in the first circle, the reason being that only in this manner would proper comparison between areas in the two States under the same crop be made conveniently. It follows, therefore, that when only one circle is drawn to represent components of one magnitude, the sectors in it should be arranged in some order—ascending or descending—and, when two or more than two circles are drawn, the sectors should have the same order in all the circles.

It is, no doubt, theoretically possible to use circles and sectors for showing the distribution of different incomes over different items of expenditure; but in practice it is not done for two reasons. Firstly, it is easier to draw sub-divided rectangles than to draw sub-divided circles, for the calculation of angles of sectors involves considerable labour. Secondly, the heights of sub-divisions of a rectangle are measurable on the scale showing percentages, and are, therefore, comparable directly in percentages; while the sectors of a circle are comparable directly in terms of angles and only indirectly in terms of percentages. Therefore, rectangular diagram is to be preferred for presenting family budgets to circular diagram. Circle divided into sectors, however, is a good diagram for showing the distribution of world population into various continents or of world area into areas for different continents.

Three Dimensional Diagrams—Cubes.

When quantities which have the ratio of 1:1000 are to be diagrammatically represented, even squares and circles, to say nothing of bars, fail to serve the purpose, for if circles or squares are drawn in such a case their radii sides will have to be in ratio of 1:32, which dimensions are difficult to show on the same scale. In cases like this, surface diagrams, such as

squares and circles, are abandoned in favour of volume diagrams or three dimensional diagrams, such as cones, blocks, spheres, cubes, etc. Of these, cubes are the easiest to draw and are generally used. The sides of cubes are equal to the cube-root of the data to be presented. The sides of two cubes representing 1 and 1000 will be in the proportion of $\sqrt[3]{1} : \sqrt[3]{1000}$, or 1 : 10. Cubes should be used only in those cases in which the data cannot be adequately presented through bar or surface diagrams.

Table 42 gives the production of tea in certain provinces of British India and Coorg before Independence. The quantities

Table 42. *Production of Tea in some Provinces of India and Coorg in* 1939.

Provinces	Production in '000 lbs.	Production (Reduced figures	Cube roots	Side of cube in inches
Bengal (Undivided)	1,12,290	864	9·6	1·9
Madras ..	38,872	299	6·7	1·3
Punjab (Undivided)	2,807	22	2·8	·6
Bihar ..	1,335	10	2·2	·4
Coorg ..	130	1	1·0	·2

giving the production are very great, and therefore they are reduced to smaller ones by dividing each of them by 130, the lowest quantity in the table. Cube-roots of these reduced figures are then found out, but since the values of the cube-roots are sufficiently large, each of them is divided by 5 and the resulting figures, given in the last column of the table, constitute the sides of the different cubes.

The construction of a cube requires explanation. Suppose the side of a cube is 1·9 inches, as it is in the case of Bengal. First, a square with its side equal to 1·9 inches is drawn. Then another squares of the same size is placed behind it in such a way that a quarter of each of the two squares just covers each other. This being done, the corresponding corners of the two squares are joined and the construction of the cube is complete. In Figure 11, the method of constructing the cube is given separately. Figs. I and II show the position of two squares ABCD and *abcd* having equal lengths of their sides, *abcd* is placed over ABCD in Figure III in such a manner that *ad* and *dc* are bisected by AB and BC; *ad* and *dc* are then rubbed off, and the corresponding points A and *a*, B and *b*, and C and *c* are joined. The sides to be rubbed off are shown by dotted lines in figure III. ADC*cba* is the required cube.

Production of Tea in 1939 in some Provinces of India and Coorg

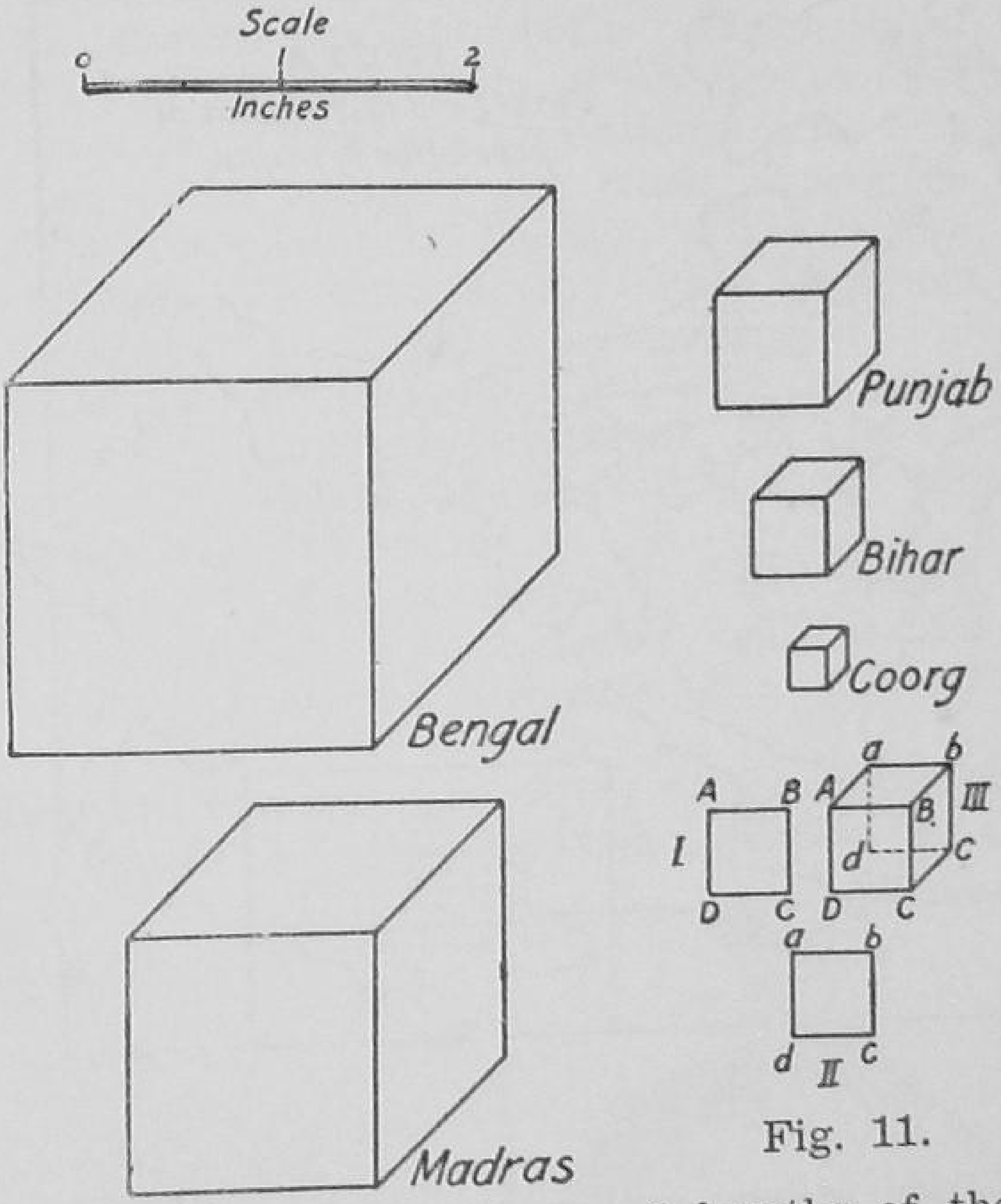


Fig. 11.

In this manner all cubes have been drawn in Figure 11 to represent the data given in table 42. They are all drawn on the same scale and are, therefore, comparable with one another.

Pictograms—Map and Pictures.

The device of pictures is being profusely used now for comparing statistical data. One comes across pictorial representation of facts very often in various exhibitions, propaganda posters, etc. They are, in fact, very effective for propa-

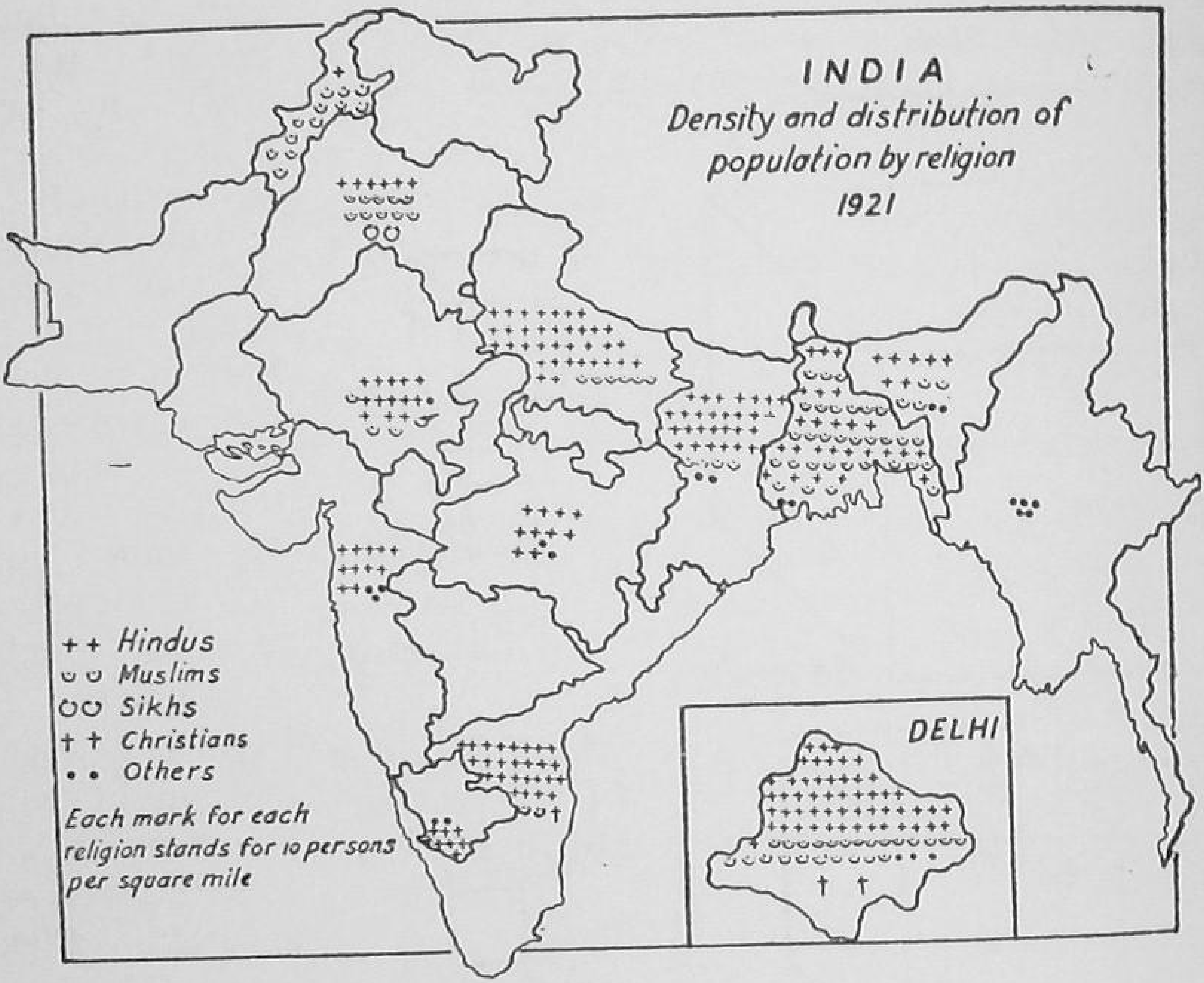


Fig. 12.

ganda purposes. They present dull masses of figures in interesting and attractive manner through objects of daily observation. The image of the entire data is fixed in the mind of the observer by a mere glance at the picture. Relationship between figures and their comparison can be studied through pictograms much more easily than by studying huge mass of numerical data.

Figure 12 shows the density and distribution of population in India prior to Partition by religion. It shows two facts at one and the same time, *viz.*, the density of population

in different provinces, and the proportions of people belonging to the prominent religions of the country. From a study of the map it will be found that the total number of marks for different religions for the undivided Punjab is 18 which means that the density of population in the province is 180 persons per square mile. Out of these 18 marks, 6 stand for Hindus, 10 for Muslims and 2 for Sikhs. Hence out of 180 persons per square mile in the Punjab, 60 are Hindus, 100 are Muslims and 20 are Sikhs, Christians and others being insignificant. It will be seen in the map that the mark chosen for each religion in itself corresponds with the sign of the religion concerned, *e.g.*, the Swastika for the Hindus, and the Moon and the Star for the Muslims. This enables us to compare at a glance the proportions of different religions along with the density of population. Instead of using religious signs, we could have as well put down typical human figures representing people belonging to different religions.

Figure 13 contains pictures of two money bags, one showing the expenditure on industries and the other that on police by the 11 provincial governments in 1940-41 in India before Partition. The expenditure are respectively Rs. 115 lakhs and 1120 lakhs. The money bags are not drawn at random.

Pictogram showing expenditure of the eleven provincial governments in India in 1940-41 on Industries and Police.

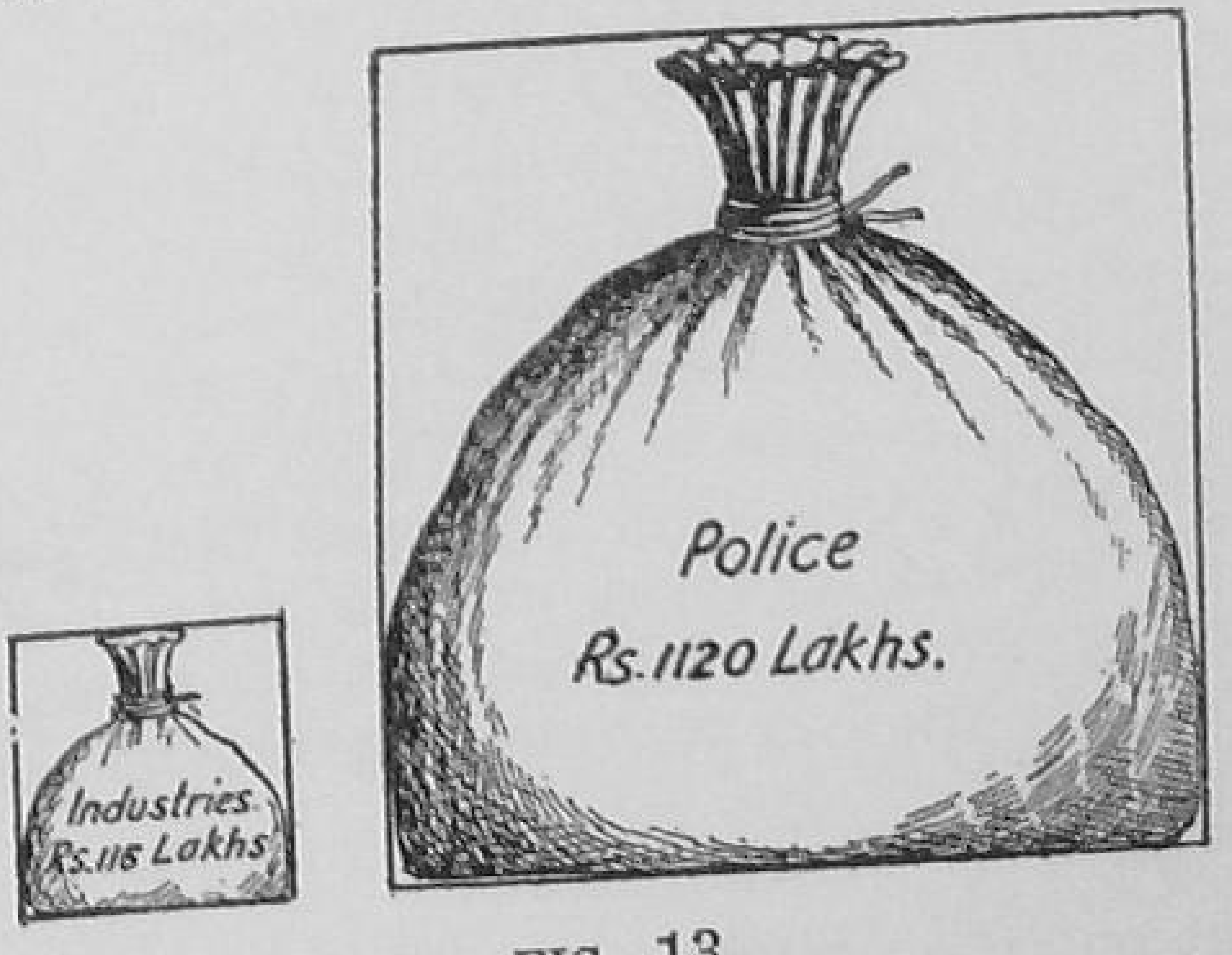


FIG. 13.

They are enclosed in squares whose areas are in the proportion of 115:1120, or whose sides are in the proportion of $\sqrt{115}:\sqrt{1120}$, or 10·72:33·47 or 1:3 approximately. A mere glance at the bags enables a comparison of the amounts spent on the heads of industries and police by the provincial governments in undivided India in 1940-41.

The money bags could have been placed inside bars or rectangles whose width may have been equal and heights in proportion to the amounts. But, then, the shapes of the bags would have looked unnatural, and instead of being attractive the pictures would have become repulsive. Great caution is, therefore, necessary in presenting statistical data in pictorial form. Pictures cannot be used indiscriminately. Serious thought must be devoted before using them, lest they might look unnatural or ridiculous, or mis-represent the data. Pictures are usually enclosed in squares, circles or rectangles for the reason that it is not possible to calculate the surface area of pictures. The pictures of the bags, for example, are very irregular so that their areas cannot be easily determined. They have, therefore, been put into squares.

General Remarks.

Several devices of representing statistical data diagrammatically have been considered above. Forethought with regard to the suitability of a particular form in a given case and practice in drawing diagrams are essential factors. Bars and circles, it will be found by experience, are the easiest to draw and suitable for general use. A particular case, however, may necessitate a square, a rectangle or a cube. Attention, in drawing diagrams, should always be fixed upon their neatness and on the precision with which they represent facts. As has been already said, continuous changes over a long period of time are best shown by graphs, and not by diagrams. Diagrams should, therefore, be used in representing spatial series.

EXERCISES

(1) Point out the usefulness of diagrammatic representation of facts, and explain the construction of any one of the different forms of diagrams you know.

(B. Com., Alld., 1945).

(2) What precautions are necessary in drawing a good diagram ? What mistakes are generally found in diagrams ? How would you avoid them ?

(3) Show with the help of a few examples that diagrams can be wrongly used.

(4) Write short notes on:
Bar-diagram, pie-diagram, three-diamensional diagram, pictogram.

(5) What kind of statistical data are best represented by diagrams ? Illustrate your answer with examples.
(B. Com., Agra, 1937).

(6) Illustrate the following by suitable diagrams:—

(*a*) In 1931 twelve seers of wheat could be had for one rupee, while in 1943 a rupee purchased only two-and-a-half-seers of wheat.

(*b*) 120 out of every 1,000 of the population of India were literate in 1941, as against 95 ten years past.

(*c*)

		I	II
Price per unit of a commodity	..	Rs. 10	Rs. 12
Quantity sold ..	..	Rs. 20	Rs. 24
Value of raw materials used	..	Rs. 100	Rs. 120
Other expenses of production	..	Rs. 60	Rs. 96
Profit	..	Rs. 40	Rs. 72

(7) The First Five Year Plan (Draft Outline) of the Planning Commission gives the following distribution of cost of development for U.P. and West Bengal:

Head		Cost of Development 1951-52 to 1955-56 (Lakhs of Rupees) *U.P.*	*West Bengal*
Agricultural & Rural Development	..	2588·51	1035·11
Major Irrigation & Power Projects	..	2877·00	1595·51
Industry	..	572·58	116·61
Transport	..	642·23	1557·60
Social Sciences ..	..	2431·72	2554·66
Miscellaneous ..	..	Nil	Nil
		9112·04	6877·49

Draw sub-divided pie-diagram to compare the cost of development under each head in the two States.

(M.A., Pb., Apr., 1952).

(8) (*i*) Diagrams are meant for a rapid view of the relation of different data and their comparison.—Discuss.

Draw a 'Bar' or 'Pie' diagram to represent the following data:—

Output and Cost of Production of Coal

	1924	1928
Cost per ton disposed commercially:	Rs.	Rs.
Wages	12·74	7·95
Other Costs	5·46	4·51
Royalties	0·56	0·50
Total	18·76	12·96
Sale Proceeds per ton ..	19·91	12·16
Profit (+) or loss (−) per ton	1·15	−0·80

(M.A., Agra, 1940).

(*ii*) Represent the above by sub-divided bars drawn on percentage basis.

(M.A., Pb., Apr., 1951).

(9) Represent the following by sub-divided bars drawn on a percentage basis:—

Cost, Proceeds, Profit or Loss per Chair during 1938, 1939 *and* 1940.

Particulars		1938 Rs.	1939 Rs.	1940 Rs.
Cost per Chair—				
(i) Wages	..	4·5	7·5	10·5
(ii) Other Costs	..	3·0	5·1	7·0
(iii) Polishing	..	1·5	2·4	3·5
Total Cost	..	9·0	15·0	21·0
Proceeds per chair	..	10·0	15·0	20·0
Profit (+)/Loss (−) per chair		(+)1·0	..	(−)1·0

(B. Com., Alld., 1948).

(10) The following table gives the national income of the Indian Union by industrial origin in 1948-49. Represent the data by a suitable diagram.

Item		Net Output (Rs. 100 crores)	Percentage
Agriculture ..	..	41·5	47·6
Mining, manufacturing and hand-trades ..	..	15·0	17·2
Commerce, transport and communications ..	..	17·0	19·5
Other services ..	..	13·8	15·9
Net domestic product at factor cost ..	..	87·3	100·2
Net earned income from abroad		(−)0·2	(−)0·2
National income ..	..	87·1	100·0

(11) *Value of the imports of glass and glassware into India from different countries—during the year 1931-32.*

Japan	.. ..	42 lakhs of rupees.
Czechoslovakia	.. ..	23 ,, ,,
Germany	.. ..	20 ,, ,,
U. K.	.. ..	13 ,, ,,
Belgium	.. ..	13 ,, ,,
Other countries	.. ..	11 ,, ,,

Represent the above figures by suitable diagrams.

(B. Com., Alld., 1933).

(12) Draw a simple diagram to represent the following statistics relating to the area under different crops in British India in 1933-34, and write a brief note on the given data:—

Crop.	Million Acres.
Rice	80·3
Wheat	27·6
Jowar	21·4
Other food crops	88·2
Oil-seeds	17·8
Cottons	14·5
Other fibres	3·1
Fodder crops	10·2
Other non-food crops	3·9

(B. Com., Cal., 1937).

(13) Represent the following data by a suitable diagram:

Gross Earnings on Class I Railways in India.

(Lakhs of rupees)

Particulars		April 1951	April 1952
Total Gross Earnings	.. ..	23,14	24,78
Passenger Earnings	.. ..	9,73	8,87
Other Coaching Earnings	.. ..	1,40	1,56
Total Goods Earnings	.. ..	11,68	14,02
Sundries Earnings	.. ..	33	33

(14) Draw a suitable diagram to represent the following data:

Exports and Imports of Cotton Piece-Goods—India.

('000 yards)

Year	Exports by Sea	Imports by Sea
1948	7,250	648
1949	9,000	1,743
1950	21,500	125
1951	15,115	175

(The above figures relate to some selected ports and do not necessarily show all-India totals).

(15) Utilize the following data to present diagrammatically the *relative* increase in note circulation towards the end of 1945 in the different countries:—

Notes in Circulation.

(In millions of national currency unit)

Country	1939	End of 1945
Canada ..	233	1,129
U. S. A. ..	7,598	28,507
U. K.	555	1,380
Australia ..	57	200
India	2,245	12,109

(M. Com., Alld., 1948).

(16) Draw suitable diagrams to represent the following information:—

Factory.	Wages. Rs.	Materials. Rs.	Profits. Rs.	Units produced.
A	2,000	3,000	1,000	1,000
B	1,400	2,400	1,000	800

Show also the cost and profit per unit.

(B. Com., Agra, 1943).
(B. Com., Alld., 1952).

(17) The following table gives the birth rates and death rates of a few countries of the world during the year 1931.—

Country	Birth rate	Death rate
Egypt	44	27
Canada	24	11
U. S. A.	19	12
India	33	24
Japan	32	19
Germany	16	11
France	18	16
Irish Free State ..	20	14
United Kingdom ..	16	12
Soviet Russia ..	40	18
Australia	20	9
New Zealand ..	18	8
Palestine	53	23
Sweden	15	12
Norway	17	11

Represent the above figures by a suitable diagram.

(B. Com., Luck., 1938).

(18) The following table gives the details of monthly expenditure of three families:—

Items of Expenditure	Family X	Family Y	Family Z
	Rs.	Rs.	Rs.
Food	24	60	180
Clothing ..	4	14	70
House-rent ..	4	16	80
Education ..	3	6	24
Litigation ..	2	10	80
Conventional needs	1	6	120
Miscellaneous ..	2	8	46

Represent the above figures by a suitable diagram. Which family is spending most wisely?

(M. Com., Alld., 1950).

(19) The following table gives the details of the cost of construction of a house in Allahabad:—

	Rs.
Land	4,500
Labour	2,500
Bricks	2,000
Iron	1,800
Timber	1,500
Cement	800
Lime	800
Stone	600
Sand	200
Other things	1,300

Represent the above figures by a suitable diagram.

(B. Com., Alld., 1941).

(20) Show the details of monthly expenditure of two families given below by means of two-dimensional diagrams:

Items of expenditure	Family A Income Rs. 500 p.m.	Family B Income Rs. 400 p.m.
	Rs.	Rs.
Food ..	140	120
Clothing ..	80	80
House rent ..	100	60
Education ..	30	40
Fuel & lighting ..	40	20
Miscellaneous ..	40	40

(M.A., Pb., Sept., 1952).

(21) The following table gives in arbitrary units the cost of production of a factory in biennial averages:—

	1918-19	1920-21	1922-23	1924-25	1926-27	1928-29	1930-31	1932-33	1934-35
Material ..	37	25	35	36	35	38	22	17	26
Labour ..	10	8	11	11	11	12	7	5	8
Overhead ..	14	10	15	16	17	20	12	9	12
Total ..	61	43	61	63	63	70	41	31	46

Draw graphs of the different components of costs as percentage of total cost, and comment on the data.

(M.A., Patna, 1944).

(22) Illustrate the following by suitable diagram:—

Production of Cotton in Egypt.

1911—12	..	..	3,111,000 tons
1914—15	..	..	6,450,000 tons
1918—19	..	..	10,873,000 tons

(23) Following are the figures of the population of the various countries of the world and of total world population in 1931:—

Country			Population (000's omitted)
China ..	..	..	411,770
India ..	..	..	352,370
U. S. S. R.	..	..	161,000
U. S. A.	..	..	124,070
Germany	..	..	64,776

Country			Population (000's omitted)
Japan ..	..	..	64,700
U. K. ..	..	..	46,077
France	..	..	41,860
Italy ..	..	..	41,100
Others	..	..	705,077
World	..	..	2,012,800

Represent the above data by a circular diagram divided into sectors.

(B. Com., Alld., 1949).

(24) Diagrammatically compare the following statistics of textiles production and imports in India. What conclusions do you draw from the given figures ?

	In Crores of yards	
	1913-14	1938-39
Mill Production	116·4	426·9
Hand-loom Production	106·8	192·0
Imports	319·7	64·7

(B. Com., Alld., 1946).

(25) Compare with the help of a suitable diagram the Time Deposits of Scheduled Banks in India with their Demand Deposits from the data given below:

	(Lakhs of rupees)	
Year	Demand Deposits	Time Deposits
1949	609,72	278,19
1950	594,30	282,87
1951	604,07	290,58

(26) Illustrate the following data diagrammatically:—

Area in (ooo) *square Kilometers of the continents of the world.*

Continent	Area
Asia	41,900
S. America	40,687
Africa	29,946
N. America	19,653
Europe	11,426
Oceania	8,550
Others	2,764
World	154,926

(27) Represent the following figures diagrammatically:

City	Females per 1,000 males in 1951
Calcutta	571
Madras	914
Nagpur	1,006
Lucknow	786
Bombay	597
Delhi	772
Rajahmundry	1,022

(28) Show diagrammatically the balance of trade in cotton piece-goods from the following data.

	Exports million yds.	Imports million yds.
1939-40	221·3	579·1
1940-41	390·1	447·0
1941-42	779·4	181·5

Also show in separate diagrams

(*a*) the proportion of exports and imports to total trade in 1941-42.

(*b*) the proportion of exports to imports in 1939-40.

(29) Represent the following by a suitable diagram:

Principle Heads of Revenue.	1938-39 Lacs of Rs.	1939-40 Lacs of Rs.
Custom	4050	4588
Central Excise Duties ..	868	652
Corporation Tax ..	204	238
Taxes on Income ..	1374	1420
Salt	812	1080
Opium	50	46
Other Heads	112	130

(B. Com., Nagpur, 1943).

(30) Represent the following data diagrammatically:—

('000 heads)

Country	Horses	Mules	Asses
U.S.A. ..	10,629	4,163	48
Iran ..	305	52	1,180
China ..	4,080	3,624	9,018
Egypt ..	47	29	1,069
Mexico ..	2,509	938	2,311
India ..	1,780	50	1,186

(31) The following table gives the population of the Part A States in India as per the 1951 Census. Draw a sub-divided pie-diagram to represent the data.

	Million Persons
Assam	9·04
Bihar	40·23
Bombay	35·96
Madhya Pradesh	21·25
Madras	57·02
Orissa	14·65
Punjab	12·64
Utter Pradesh	63·22
West Bengal	24·81

(32) Below is given the number of wagons loaded daily on class I railways in India. Show diagrammatically the proportion of wagons loaded under each class of goods to total loadings in different years:

Wagons Loaded—Daily Average

		1950	1951
Coal & coke	..	6,459	6,815
Grain & pulses	..	2,240	2,702
Oilseeds	..	474	410
Cotton	..	390	434
Miscellaneous	..	19,128	19,848
TOTAL	..	28,691	30,209

(33) The table below gives the percentages of total expenditure on consumption of certain commodities in six zones of India in 1949-50 according to General Report No. 1 on the First Round of the National Sample Survey. Show the data by a suitable diagram.

Population Zones (Rural Areas only)	Percentage of total expenditure on.		
	Food items	Ceremonials	Others
North	69·65	6·56	23·79
East	71·01	5·18	23·81
South	67·37	4·17	28·46
West	54·01	15·58	30·41
Central	62·79	8·23	28·98
North-west	67·94	6·47	25·59
All-India	66·31	7·21	26·48

(34) Below are given the amounts of outlay during 1951-1956 proposed in the final Five Year Plan and the Draft outline of the Plan.

Outlay during 1951-56.

(Crores of Rs.)

Item	Final Five Year Plan	Draft Outline of the Plan
Agriculture & community development ..	360·43	191·69
Irrigation & Power ..	561·41	450·36
Transport & Communications	497·10	368·12
Industry	173·04	100·99
Social Services ..	339·81	254·22
Rehabilitation	85·00	79·00
Miscellaneous	51·99	28·54
Total ..	2068·78	1492·92

Compare diagrammatically the percentages of outlay on different items to total outlay in the two versions of the Plan.

(35) Utilize the following data to show diagrammatically the revenue, expenditure and surplus or deficit of the Government of India during each of the three years ending 1953-54:

(In lakhs of Rs.)

	1951-52 Accounts	1952-53 Revised	1953-54 Budget
Revenue	5,15,36	4,18,64	4,39,26
Expenditure	3,74,06	4,22,43	4,38,81

CHAPTER XIII

GRAPHICAL PRESENTATION

Diagrams and maps discussed in the preceding chapter are particularly suited to the comparison of variables spread over different places or different heads at the same time, that is, they are useful for representing spatial series. For illustrating series spread over a period of time and also for illustrating frequency distributions, graphical methods are made use of. In this chapter, therefore, we shall discuss

(1) Graphs of Time Series showing continuous changes, and

(2) Graphs of Frequency Distributions.

Diagrammatic and Graphic Presentations Contrasted.

In the diagrammatic method bars, rectangles, circles etc. stand for quantities individually or in groups. In the graphical method quantities are not *represented* by one or more dimensional figures, but are *located* on a surface with respect to two or more dimensions, for which purpose a system of rectangular co-ordinates, such as that given on the next page is used.

Two straight lines x x' and y y' intersecting each other at right angles are drawn. The horizontal line is the x-axis or *abscissa* and the vertical line the y axis or *ordinate*. The junction of the axes at o is known as the point of origin or zero. Distances measured towards the right or upwards from the origin are reckoned as positive, and those measured towards the left or downwards as negative. All points in the plane, *i.e.,* the four quadrants into which the graph is divided, are located by reference to two co-ordinates drawn parallel to the axes.

The scales of measurement on the axes may be chosen at convenience. There is no necessary connection between the x

A System of Rectangular Co-ordinates.

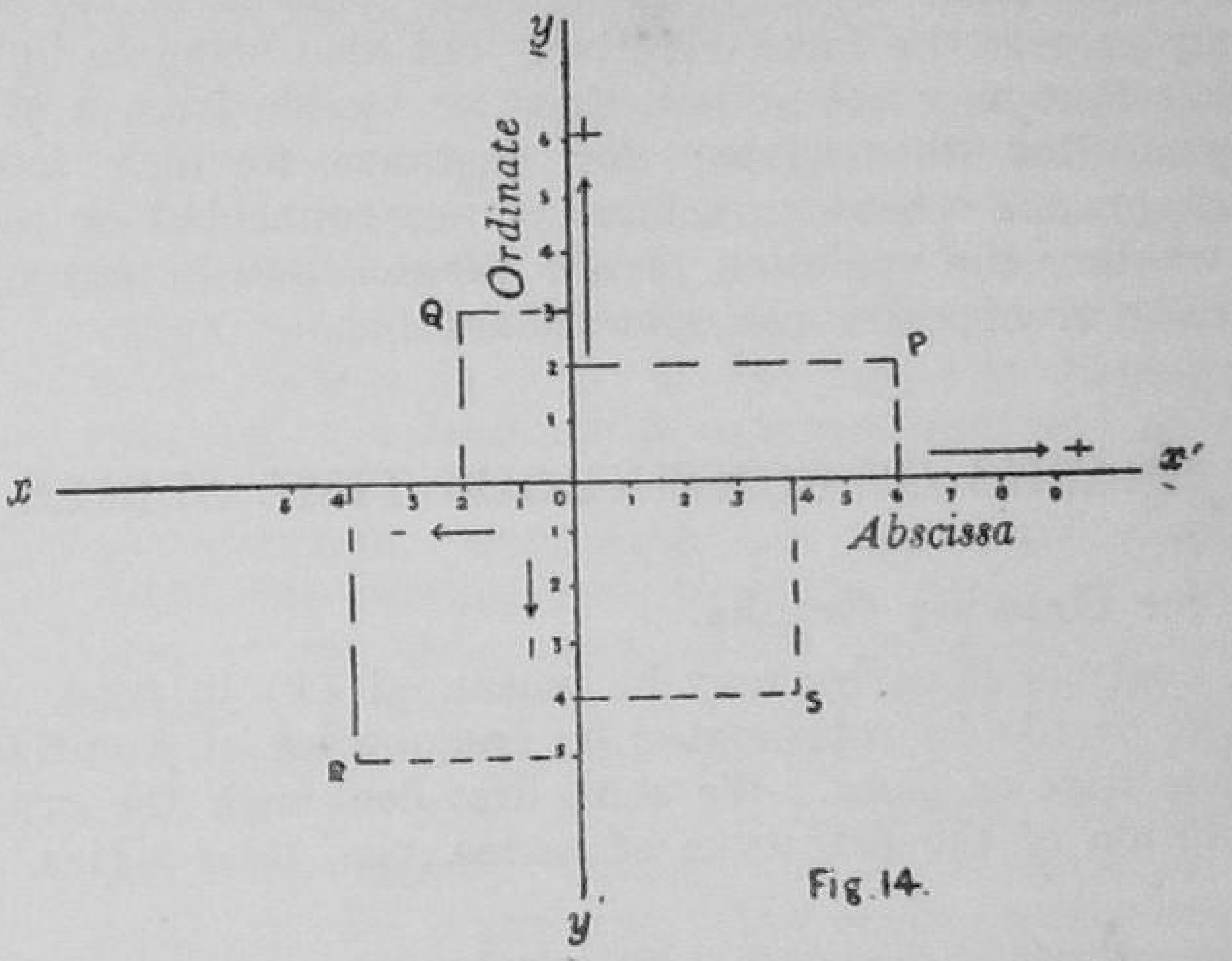


Fig. 14.

and the y scales. In Figure 14, the scales of x and y axes are equal and the points P, Q, R, S have been located as follows:—

Point	x	y
P	+6	+2
Q	−2	+3
R	−4	−5
S	+4	−4

The co-ordinates on the points are indicated by dotted lines. But the dotted lines need not be made visible; only the point should be shown. In practice, graph papers, which have a net-work of fine lines, are used for graphical presentation. By using them, the necessity of drawing lines is dispensed with.

It is evident from the above that to *locate* quantities in a plane bounded by the x and the y axes in the manner illustrated in Figure 14 is not the same thing as to *represent* them by bars, squares, rectangles, etc. The latter are themselves drawn proportional to the amounts which they represent.

All truly continuous series are properly presented by graphical as distinct from diagrammatic methods. Graphic methods are more powerful than diagrammatic ones in so far as they not only present the facts effectively but also bring to light new relations that may not at first sight be visible from a study of the quantities themselves; for instance, we may determine through graphs whether phenomena are connected or independent, whether the variation in one phenomenon is accompanied by similar or opposite variation in another.

GRAPHS OF CONTINUOUS TIME SERIES

Rules for Drawing Graphs.

A continuous series may be measured (1) in time, or (2) in space, or (3) be represented by frequencies of a variable at the same time or place. We shall first deal with the graphical presentation of the first type of series, *viz.*, time series.

Time series is also called **historical series**, since it stands for the numerical record of the changes in a variable during a number of successive intervals in a period of time. The first problem to be considered in drawing a graph of such series is the choice and adjustment of scales.

Choice and adjustment of scales.—A system of rectangular co-ordinates as illustrated in Figure 14 is used to illustrate time series. Time units are placed on the abscissa or x-axis, and the sizes of variable measured on the ordinate or y-axis. Since time has no zero, the horizontal scale need not begin with zero; the first time interval may be indicated near or away from the point of origin or 0. The x-axis should be divided into equal parts, each of which should represent periods of equal length. The vertical scale should begin with zero when sizes of variable are shown on it, since they are always reckoned from zero. Equal space units on the y-axis should represent equal amounts when natural scale is used, and equal rates of change when ratio scale is used. Just now, we shall be concerned with the natural scale.

What should be the proportion between the abscissa and the ordinate scales? Bowley states the problem and the way in which it should be solved as follows:—

> "It is difficult to lay down rules for the proper choice of the scales by which the figure should be plotted out. It is only the ratio between the horizontal and vertical scales that need be considered. The figure must be sufficiently small for the whole of it to be visible at once; if the figure is complicated, relating to a long series of years and varying numbers, minute accuracy must be sacrificed to this consideration. Supposing the horizontal scale decided, the vertical scale must be chosen so that the part of the line which shows the greatest rate of increase is well inclined to the vertical, which can be managed by making the scale sufficiently small; and, on the other hand, all important fluctuations must be clearly visible, for which the scale may need to be increased. Any scale which satisfies both these conditions will fulfil its purpose."[1]

Thus, the scales chosen should be such as would allow the full data to be presented on the graph, would properly show the extreme fluctuations and would clearly bring out the changes over the entire period from date to date. The two scales selected will, no doubt, depend upon the size of the paper; still it should be carefully noted that if the units occupy too much space, small changes in the size of variable will appear to be important fluctuations while, if the units occupy too little space, even large fluctuations would look unimportant. The respective scales will, it is obvious, be different for different data. **No one standard can suit all cases, yet, it is desirable as a general rule, to have the x-axis approximately $1\frac{1}{2}$ times as long as the y-axis.**

Having decided the proportion of the two scales, the ordinate scales hould be divided into units such that they would be easily comprehended in terms of the rulings of the paper

[1] Bowley, A. L., *Elements of Statistics*, 1920 ed., p. 132.

used. For instance, if the paper is ruled in fifths or tenths, ten small squares should not be made equal to such an amount as 4357. A given space should equal some multiple of ten or five, as 300, 250, 25 etc. When the scale has been divided into units, the ordinate should be labelled in terms of the scale unit, *i.e.*, at each equal space the value of it in terms of the scale unit should be put down. One should not try to fill up the scale with too many details, and the putting down of each successive frequency—or every frequency that is plotted—should be avoided.

The natural scale is thus ready. We shall deal later with ratio scales and false base line. Presently, the problem of plotting the data is taken up.

Plotting the data.—To plot the given data the method shown in Figure 14 will be followed. To plot the size of an item against a particular date, only a point placed on the ordinate concerned is enough. Thus when the whole data are plotted, there will be as many points on the plane as the number of dates. Since time is a continuous factor, these several points should be connected from date to date by continued smoothed lines, each point being simply a conventional stopping place. This continuous smoothed line is called the **curve**; it shows the probable changes at all possible intervals of the entire period to which the data relate. This curve is also given the name of **Historigram,** for it presents a historical series. This name must be distinguished from Histogram, into the construction of which the factor of time does not enter. But drawing a smooth curve requires practice and skill which everyone does not usually possess. An alternative which is commonly adopted, therefore, is that of connecting the points plotted by *straight lines.*[2] When the curve is plotted, it should be given a short but adequate title, indicating what it represents. And, if several curves are plotted, each should be differentiated either by using different inks or by adopting some such devices

---

[2]In Mathematics the term *curve* includes a *straight* line, and its equation is $y=a+bx$.

as drawing straight continuous line, dotted line, dot-and-bar line, if the same ink is used. An explanation of what the different inks used or lines drawn indicate in the historigram should be separately given in a corner of the paper as given in Figure 16, or the name of the factor represented may be written on the curve itself if it does not spoil the chart, as done in Figure 30.

GRAPHS ON NATURAL SCALE.

The purpose of graphical presentation as that of diagrammatic representation is comparison. Comparison is necessary to have a clear idea of the relationship of things in time and space. The methods by which comparison in space is rendered possible have already been considered in the preceding chapter. Here, the purpose is to consider the methods by which comparison in time is made. In such comparisons, our aim may be to study:

1. Changes of a single variable over a period of time; and,

2. Changes of two or more variables over a period of time.

These changes can be studied on the Natural Scale or "Difference" Charts and on the Ratio Scale or "Ratio" Charts.

First we take up the natural scale graphs. On it, changes of a single variable are studied through (1) Absolute historigram and (2) Index historigram. And, changes in two or more variables are also compared through (1) Absolute historigrams and (2) Index historigrams. In addition to these, we shall discuss a few other ways of drawing graphs for comparison different variables, and also the false base line.

Absolute Historigram of one variable.

When original quantities, and not index numbers, are presented graphically, the resulting figure is called 'absolute historigram' as distinguished from an 'index historigram' which

relates to an index number series. In Figure 15, amounts of treasury bills tendered in undivided India from week-ending 10th February, 1942, to week-ending 28th April, 1942, given in table 43, are shown. One small division represents on the x-axis one week, and on the y-axis Rs. 5,000,000. These scales give an outline which is neither too flat nor too angular and shows the fluctuations clearly. Since all the quantities in table 43 are positive, only the north-east quadrant of the system of rectangular co-ordinates has been utilised. Only one variable has been treated in the figure and changes in its size are comparable through the graph from week to week. Thus, the amount of treasury bills tendered increased from Rs. 30,000,000 on 17th March, 1942, to Rs. 33,000,000 on 24th March, 1942, that is, it increased by Rs. 3,000,000.

Table 43. *Treasury Bill Tenders in undivided India.*

Week-ending 1942 x		Amount tendered Rs. (000,000) y
February	10th	22
,,	17th	28
,,	24th	27
March	4th	23
,,	10th	24
,,	17th	30
,,	24th	33
,,	31st	31
April	7th	30
,,	14th	24
,,	21st	27
,,	28th	35

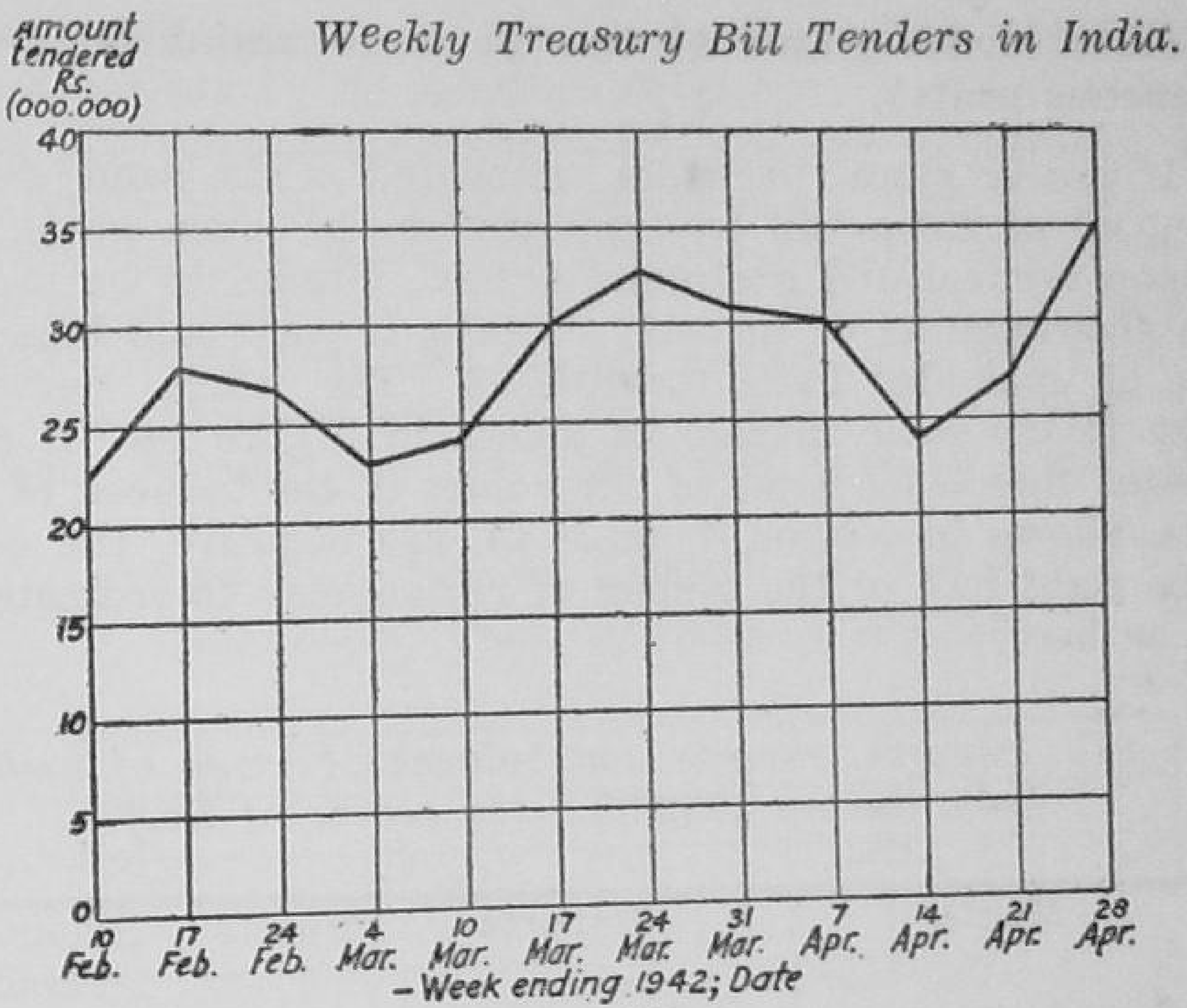


Fig. 15.

Index Historigram of one variable.

When a time series consisting of index numbers is given, it may be graphically presented in just the same manner as the series consisting of absolute or actual values. In Figure 28, an index historigram relating to the retail price of wheat in India is shown along with other curves. An important difference between index historigram and absolute historigram is that the latter shows the actual values and therefore studies the movement of absolute sizes of the variable from date to date, while the former shows the index numbers and therefore studies the percentage change on a particular date as compared with the *base* year. Thus, from Figure 28 we can study that retail price of wheat in 1877 was two per cent. more than in 1873, the base year; we cannot study the absolute amount of money by which the increment was effected.

Absolute Historigrams of two or more variables, (Homogeneous units).

If two or more variables measured in the same unit are given, all of them can be exhibited on the same graph, with *common* vertical and horizontal scales. Figure 16 on the next page shows values of monthly exports, imports and balance of trade of undivided India for 1932-33. The three curves are drawn in the same manner as shown in Figure 14. It should be noted that since some of the values in the balance of trade series, shown in column 4, table 44, are negative, the eastern or the right half of the system of rectangular co-ordinates has been utilised.

Table 44. *Imports, exports and balance of trade of undivided India during* 1932-33. (*In crores of Rs.*)

Month x	Imports y	Exports y	Balance of trade y
April	13	11	−2
May	12	10	−2
June	12	10	−2
July	11	9	−2
August	11	10	−1
September ..	11	13	+2
October	10	12	+2
November ..	11	12	+1
December ..	10	13	+3
January ..	11	12	+1
February ..	9	12	+3
March	11	13	+2

In Figure 16, comparison can be made only of the absolute amounts of exports, imports and balance of trade, since the historigrams show absolute values. It will not be possible to compare proportional changes between them during the same period.

Absolute Historigram of two or more variables, (Heterogeneous units).

If two or more variables are measured in different units, all of them can be exhibited on the same graph, but with *different* ordinate scales, the horizontal scale being common. The ordinate scales will have to be different for the simple reason that the variables are not measured in the same unit. With this difference only, the curves shall be prepared in just the same manner as they have been done in Figure 16. A

Values of Monthly Exports, Imports and Balance of Trade of India in 1932-33.

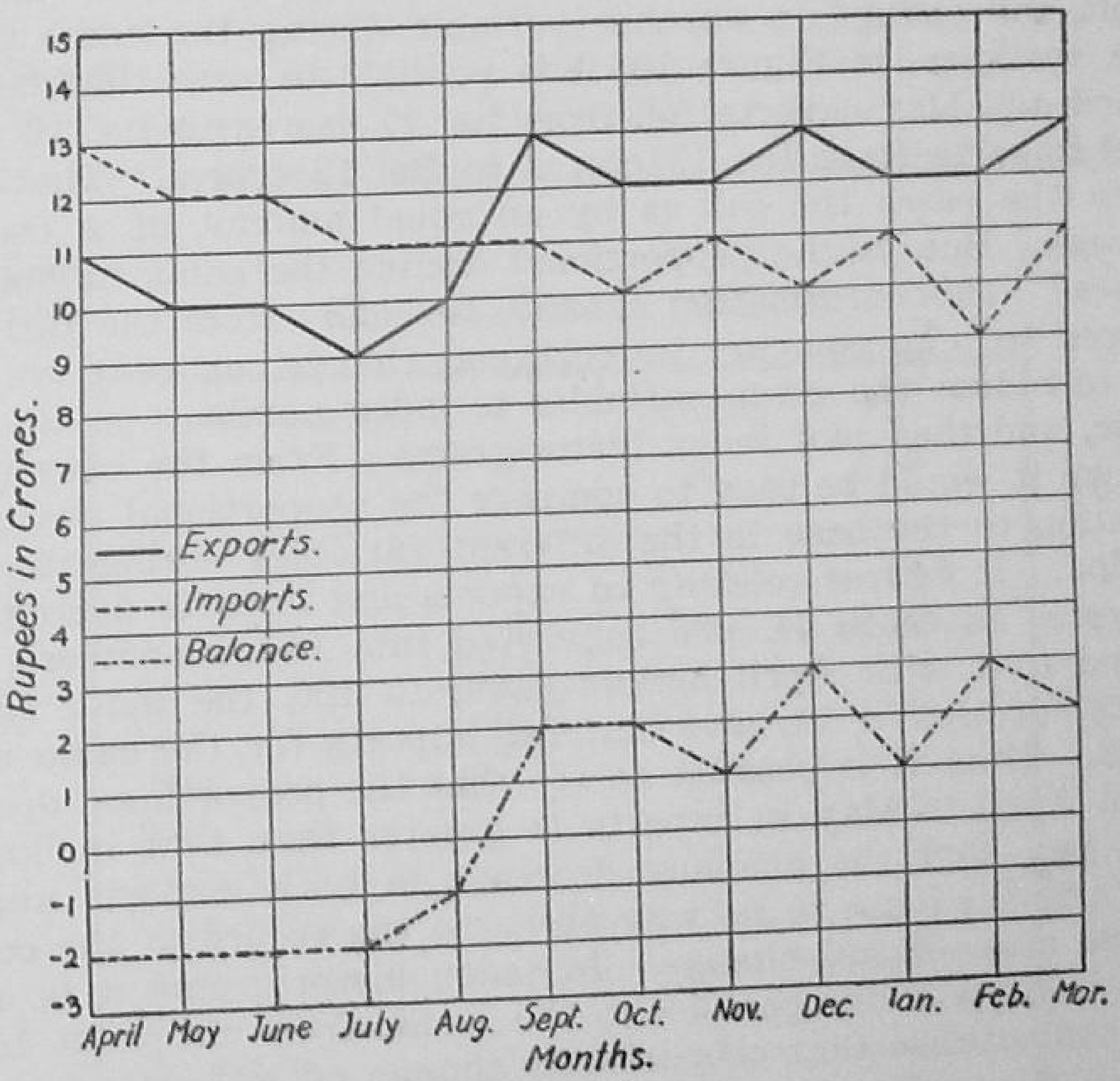


Fig. 16.

study of the changes in the absolute amounts of the same variable from one date to another will be possible, but com-

parison of changes in one variable with those in the other during the same period will not be possible. For, of two variables if one is measured in tons while the other in yards, tons will be comparable with tons and yards with yards; tons will not be comparable with yards.

Index Historigrams to compare changes of two or more variables.

In the case of absolute historigrams of two or more variables expressed in the same or different units, it has been noticed that only absolute sizes are comparable and it is not possible to study from them whether change in the value of one variable from one date to another is similar to or different from the change in another variable during the same period. For instance, in Figure 16, it is possible to note that between April and May exports fell from Rs. 11 crores to Rs. 10 crores and imports from Rs. 13 crores to Rs. 12 crores. That is, in both the cases the fall is by an equal amount of a crore of rupees. But, is the proportional decline the same in both the cases? This information cannot be had from the figure referred to. To compare proportional change, an easy way will be to reduce the given variables to index numbers on the *same base,* and then plot index historigrams. From the curves thus drawn it would be easy to compare the proportional changes in relation to the base in the different variables during the same period. If figures relating to exports and imports during May, to refer to table 44, are converted into index numbers, it is found that with April figures equal to 100, the index for exports for May is 91, and that for imports for the same month is 92. Thus, it is possible to see that the proportional decrease from April to May in exports is greater than that in imports. The fact that the absolute decrease in both, exports and imports, is the same in no way obscures the record of the comparative proportional change. In index historigrams it is advisable to draw a line parallel to the x-axis from the point 100 on the ordinate, so that proportional change on any date from the base, whose value is put down at 100, may be seen at a glance.

But, as was noticed in dealing with index historigrams of a single variable, index historigrams are no improvement on the absolute historigrams if it is desired to compare different

periods in the same series with regard to the relative changes therein. All index numbers compare the change with the base year and not between themselves. Therefore, index curves of exports and imports, with April as the base in the example under consideration, shall exhibit the comparison of proportional change with April; study of proportionate changes from August to September, or from December to January, will not be possible. The function of clearly exhibiting relative changes over periods of time is reserved for logarithmic historigrams, which we shall study later.

Methods of scale conversion for comparing changes in two or more variables.

It has been noted above that when variables are expressed in different units, or even in the same unit, they may be converted into index numbers to compare proportional changes relative to the base. Several methods of comparing the differ-

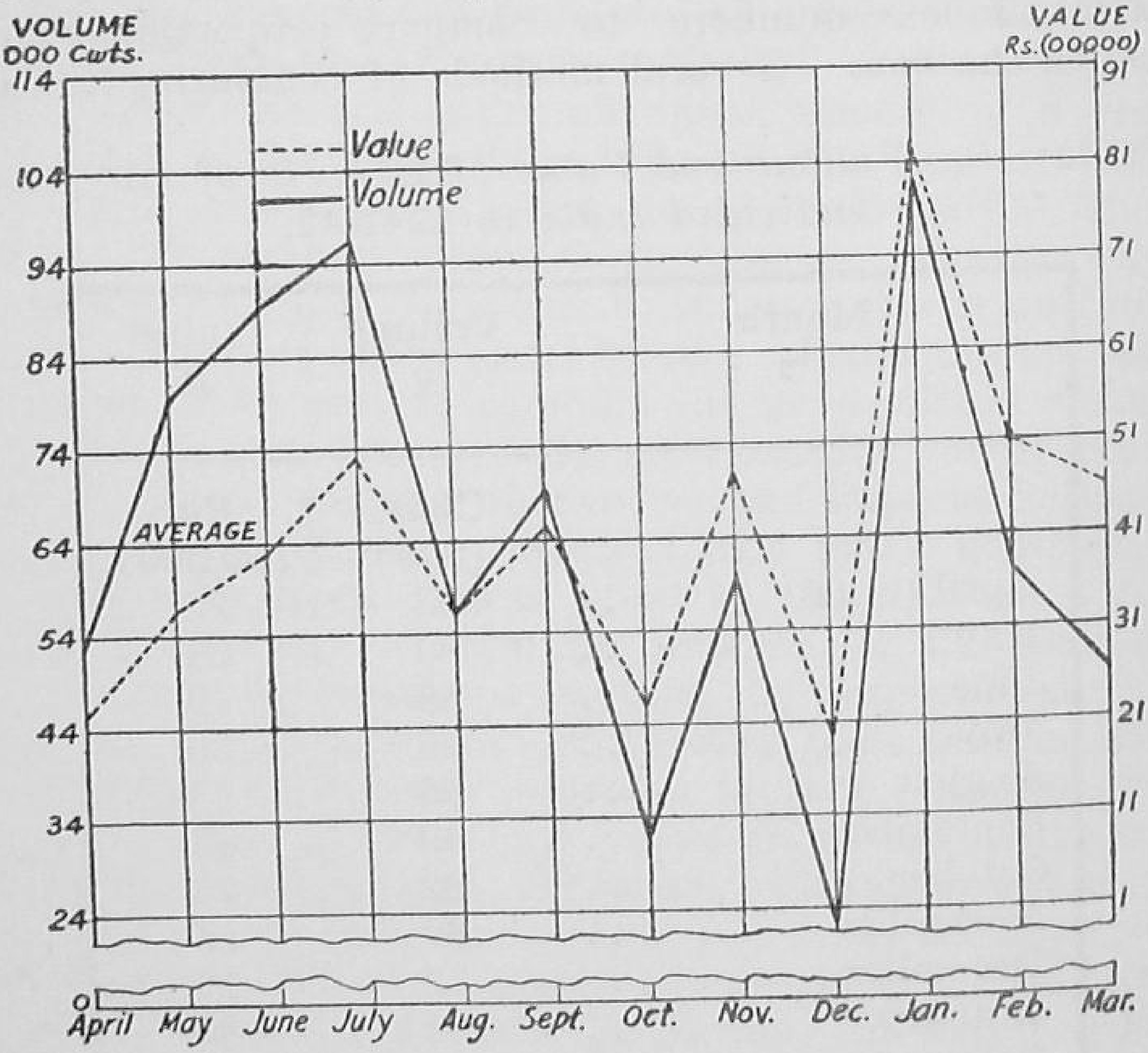
Tables 45. *Volume and Value of Exports of Lac from undivided India in* 1941-42.

Month x	Volume y	Value y
	Cwts. (000)	Rs. (00,000)
April	53	22
May	80	34
June	89	40
July	96	50
August	56	33
September ..	69	43
October	32	23
November ..	60	48
December ..	22	19
January ..	102	83
February ..	60	51
March	49	46

ences between two or more variables during a particular period are, however, available. These relate to converting one scale into terms of the other scales. Of these methods we take up one below, which is particularly suitable in a case where variables are expressed in different units.

The method is very simple. When two or more variables are given, separate scales may be chosen for different variables, but each should be made proportional to the respective averages of each. Table 45 gives the volume in cwts. and value in Rs. of lac exported from undivided India, month by month, during 1941-42. First, averages of these two series are computed which are, respectively, 64,000 tons and Rs. 41,000. These average values are plotted in Figure 17 on two separate

Volume and Value of Exports of Lac from India in 1941-42.



vertical scales in such a way that the average values of the two are represented by the same position on the vertical scales.

After the scales have been thus adjusted, the values of the variables are plotted. Each of the two curves should be read in terms of its own scale.

One difference between this method and the method of index historigrams discussed above is that in this method actual values are plotted, while in the other method values relative to the base year—that is, index numbers—are plotted. In Figure 17, it is easy to compare the two series since their averages lie on the same line.

Mixed Graphs.

Another method of presenting two or more variables expressed in different units is to draw a mixture of bar diagram and curve. Consider the following data:

Area, Production and Yield per acre of Jute in India.

Year	Area (million acres)	Production (million bales of 400 lbs. each)	Yield per acre (lbs.)
(1)	(2)	(3)	(4)
1946-47	0.54	1.32	983
1947-48	0.65	1.66	1,017
1948-49	0.84	2.07	986
1949-50	1.16	3.09	1,062
1950-51	1.45	3.30	908
1951-52	1.95	4.68	959

Ignoring the units in which area and production have been expressed, it will be noticed that the figures in Col. (2) are less than one-half of those in col. (3) so that area scale may be taken as half that of production. Thus, while one inch on vertical line stands for 1 million acres in the case of area, it stands for 2 million bales in the case of production. It is also clear that if production figures are presented by bar diagram and area figures by a curve, the curve will cut across the bars, thus giving an attractive look to the resulting Mixed Graph. This has been done in Figure 17A. The scales for the production bars and area curve are obviously unsuitable for presenting yield per acre, which has been shown separately as an inset in the Figure.

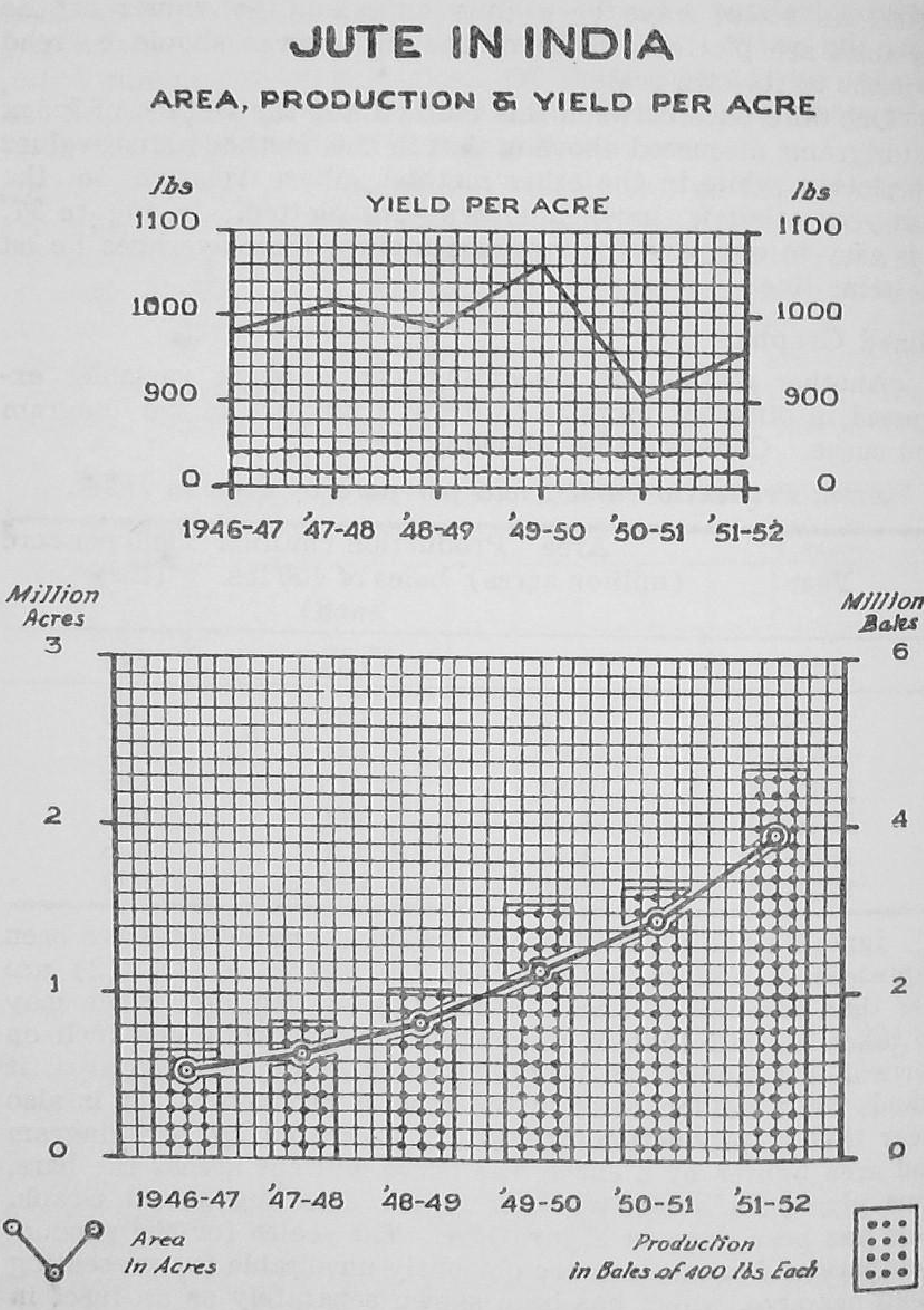

JUTE IN INDIA
AREA, PRODUCTION & YIELD PER ACRE
lbs
YIELD PER ACRE
lbs
1100
1000
900
0
1946-47 '47-48 '48-49 '49-50 '50-51 '51-52
Million Acres
Million Bales
3
2
1
0
6
4
2
0
1946-47 '47-48 '48-49 '49-50 '50-51 '51-52
Area in Acres
Production in Bales of 400 lbs Each

Fig. 17A.

Area under jute in Figure 17A shows a progressive increase during the period under study; but increase in production is not commensurate with increase in area in all the years (*cf.*, for example, the year 1950-51 with 1949-50). The inset shows that yield per acre was the lowest in 1950-51, but since the production in that year was the highest during 1946-47 to 1950-51, it is easy to attribute this high production to increase in area. The latter is also true of 1951-52, when the yield per acre was lower than that in any of the years 1946-47 to 1949-50 but the production was the highest on record since 1946-47. On the other hand, the increase in production in 1949-50 could be attributed more to rise in yield per acre than to that in area.

False Base Line.

In Figures 15 and 16, the scales on the ordinate beginning from zero are continuous, while in Figure 17, the scale is broken or discontinuous. In Figure 17 we have used the false base line.

In cases where the fluctuations are small relatively to the size of the variable but the fluctuations are to be shown clearly and in those cases where adjustment of scales is not otherwise possible (as in Fig. 17), instead of showing the entire vertical scale from zero to the highest value involved only as much of it is shown as is just sufficient for the purpose. That portion of the scale which lies between zero and the smallest value of the variable is omitted. This has been done in Figure 17 and is further illustrated in Figures 18 which shows the index numbers of prices of Government Securities during 1937-42. The range of the series represented in the latter Figure is (123−104)=19, and the fluctuations are small. To amplify these fluctuations, the vertical scale shown is only between the limits 103 and 124. On this wide vertical scale the fluctuating movement in the price of Government Securities is brought into prominence. If a continuous vertical scale beginning from zero and going up to the maximum value of the index numbers were used to plot the indices of the prices of Government Securities, a slightly waving line would have resulted and the fluctuations would have been quite inconspicuous. Similarly, in Figure 17 adjustment of the scales would not have been possible without resorting to the

false base line. The fact that a false base line has been taken and the vertical scale is not shown in its entirety, should always

Index numbers of prices of Government Securities during 1937-42.

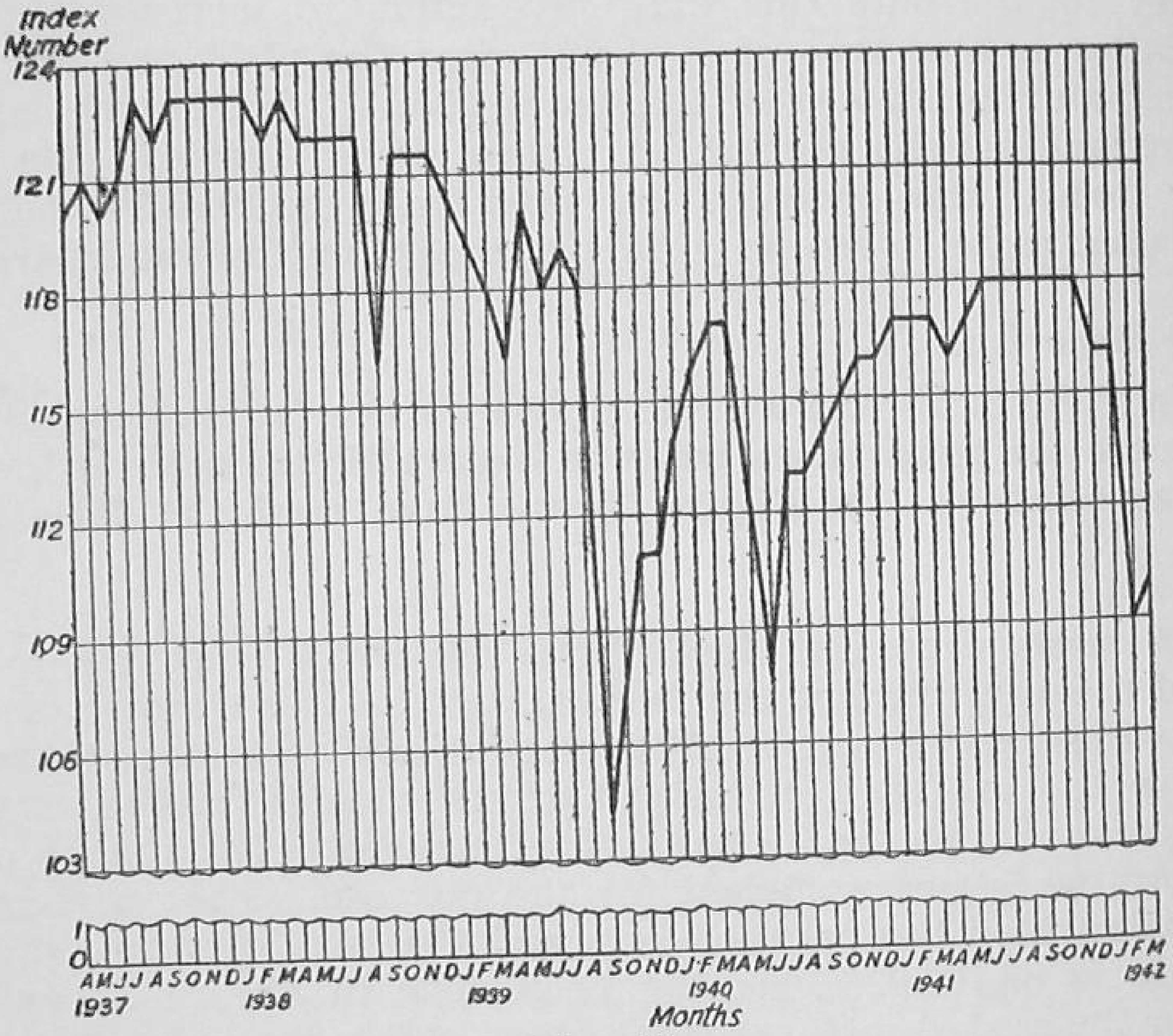


Fig. 18.

be made clear on the graph by the double saw-tooth line, as shown in Figures 17, 17A, 18 and 18A, or by some other device.

False base line should be used in rare and exceptional cases. It is always safe to show the zero line, for a correct study of the proportional changes is possible only when the zero of the ordinate in the graph is shown and the height of every point on the curve is shown in full. For example, the proportion between two numbers, 100 and 400, if expressed in full above the zero line is 1: 4. But if the vertical scale from 0 to 50 is omitted, each of the two figures will be reduced by 50. If the horizontal

line passes through 50, the respective magnitudes of these two numbers above the horizontal line will be 50 and 350 whose ratio will not be 1: 4 but 1: 7. Similarly, if the vertical scale is omitted from 0 to 70, the two quantities 100 and 400 will be reduced to 30 and 330 respectively above the base line passing through 70 of the vertical scale. Their ratio will no longer be 1: 4 but would be changed to 1: 11. Thus, wrong impressions might be created in regard to proportions if a portion of the vertical scale is omitted, *i.e.,* if a false base line is taken. This would, however, not happen when the significance of the false base line is kept in view.

Not infrequently people resort to false base line either to economise space or for want of space. Thereby they unconsciously run into the error of making fluctuations look larger than they really are. Sometimes this is done deliberately too. Therefore, while studying curves and commenting on them this fact must not be ignored.

Zone Curve.

At times it may be necessary to present graphically the maxima and minima of a series of data. This can be done by plotting two points for each given date—one for the maximum value and the other for the minimum value—and then filling in the space between the two points for each date with vertical lines. The resulting figure would be a Zone curve. Figure 18A is a Zone curve presenting the highest and the lowest quotations of spot prices of gold in Bombay on weekly basis for the period April-December, 1952. From the curve it can be easily seen that during the given period the lowest quotation reached was Rs. 77-5-0 per tola, in the week ending the 12th December, 1952, and that on that date the highest quotation was Rs. 80-4-0. It will also be seen that in the week in which the quotation was as its peak during the given period (*viz.,* Rs. 95-14-0 in the week ending the 3rd May, 1952), the lowest quotation was just equal to the highest quotation of the previous week (*viz.,* Rs. 92). The range within which the quotations fluctuated in different weeks, and the behaviour of prices over time would also become clear from the curve.

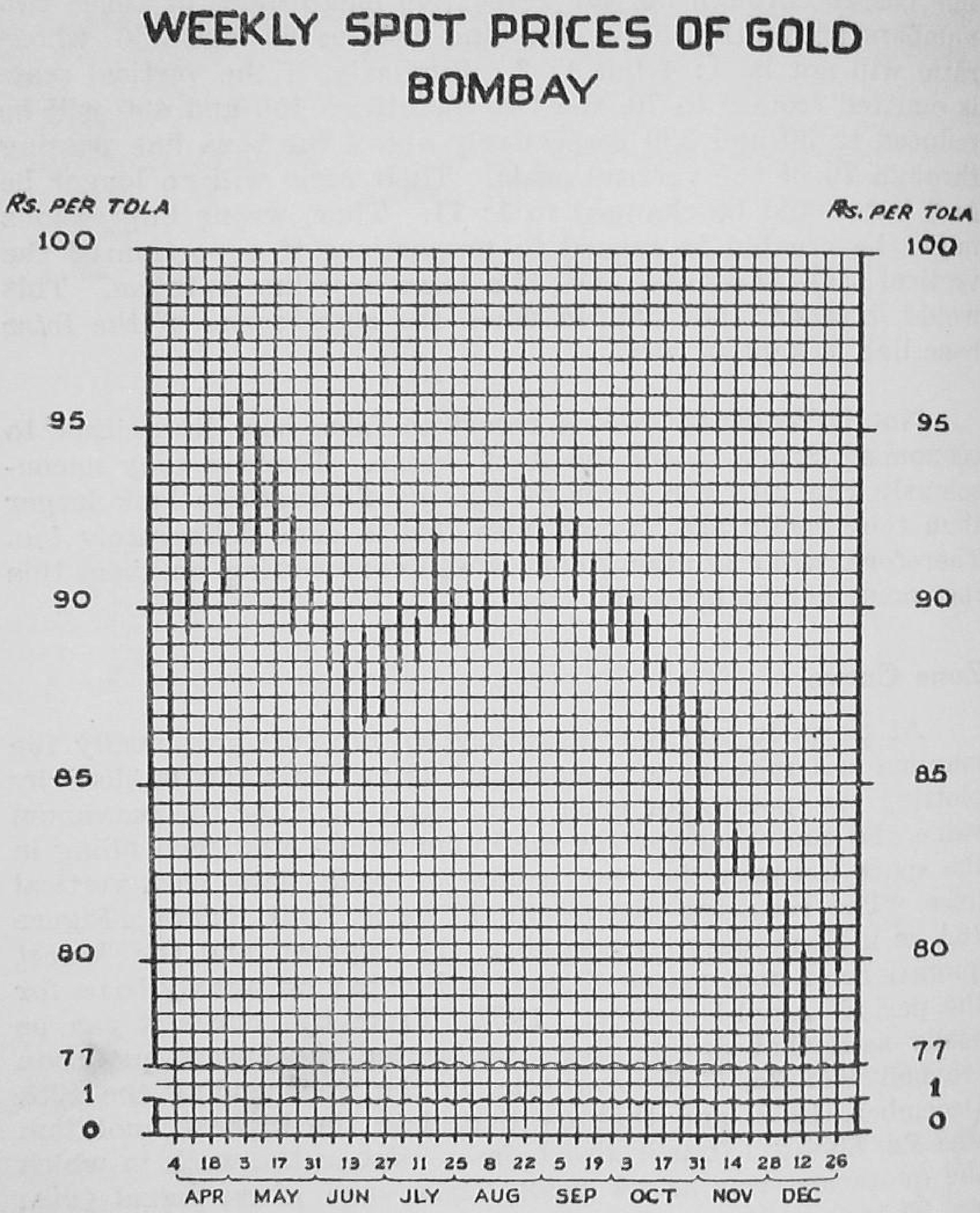

WEEKLY SPOT PRICES OF GOLD
BOMBAY
Rs. PER TOLA
Rs. PER TOLA
100
95
90
85
80
77
1
0
4 18 3 17 31 13 27 11 25 8 22 5 19 3 17 31 14 28 12 26
APR MAY JUN JLY AUG SEP OCT NOV DEC
1952

Fig. 18A.

A zone curve can be used to present interest rates, earnings, output or sales per person, salaries and wages, prices of goods, bonus, security prices, etc., over a period of time.

Band Curves.

If it is intended to show graphically the component parts of some data spread over a period of time, the device of Band curves may be utilised with advantage. The components may be plotted one above the other, thus resulting in a series of bands; the bands may then be shaded in, so that the whole space between the topmost curve and the horizontal base line is filled up. The total shaded space will show the sum of the constituents, while the different bands will show the components.

Below are given the imports on Government account into India, of cereals, classified into three components:

Imports of cereals on Government account into India.

('000 tons)

	1946	1947	1948	1949	1950	1951	1952
Wheat & wheat flour	1,344	857	1,311	2,200	1,407	3,015	2,511
Rice	326	485	867	767	353	749	722
Other cereals	579	992	663	739	365	961	631
Total	2,249	2,334	2,841	3,706	2,125	4,725	3,864

The above data are presented by means of Band curves in Figure 18B. It will be seen from the Figure that during the given period the imports of all cereals taken together were the highest in 1951, and that wheat (including flour) more than the other two constituents led to that position, the imports of wheat (including flour) in that year being even slightly higher than the combined imports of wheat (including flour) and rice in 1949. It will also be seen that the total imports were the lowest in 1950, and that this was due not so much to wheat (including flour) as to the low imports of rice and other cereals. The Figure also shows that the imports in 1949 of wheat (including flour) alone were as high as the total imports of all cereals next year ; nay even slightly higher.

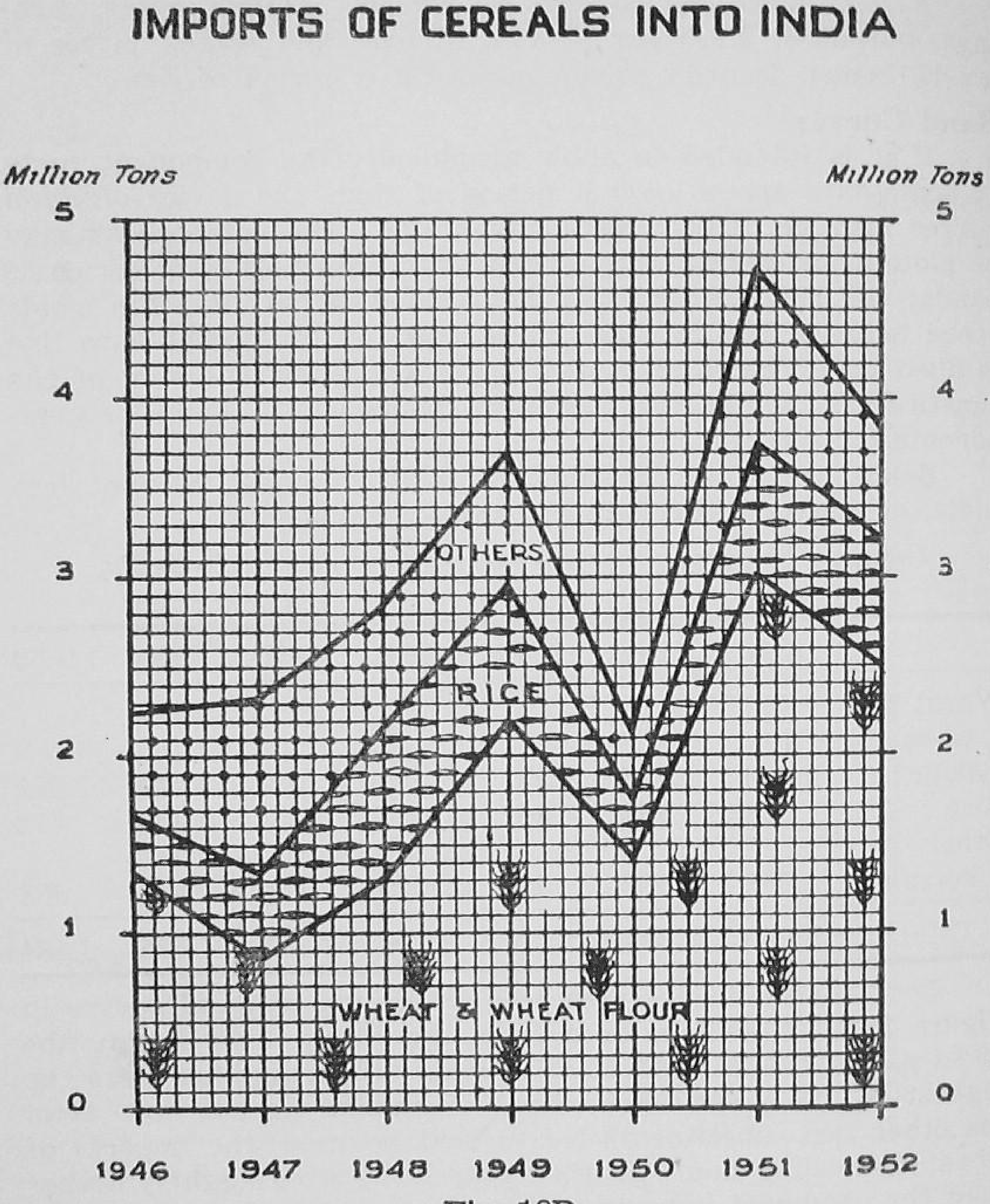

IMPORTS OF CEREALS INTO INDIA
Million Tons
Million Tons
5
4
3
2
1
0
OTHERS
RICE
WHEAT & WHEAT FLOUR
1946
1947
1948
1949
1950
1951
1952

Fig. 18B.

The Band curves in Figure 18B present the absolute quantities. The figures pertaining to the different constituents may also be converted into percentages of the total, and the percentages plotted to yield Band curves. The whole chart will then depict 100 per cent and the shaded bands will show the percentages the components bear to the whole.

Band curves can be used to show the break-up of total sales by departments, of cost and selling prices of articles, of cultivated area into different crops or groups of crops, of production of allied goods into their constituents, etc., the data in all these cases being spread over a period of time.

Zee Chart or Z Curve.

The Zee chart consists of three curves. Since the shape of the figure giving the Zee chart resembles the letter Z, it is also known as Z curve. The three curves are:—

(i) a curve of the original data,
(ii) a curve of the cumulative data,
(iii) a curve of moving *annual* totals, (annual, if the original data are monthly or weekly figures).

The table below gives monthly production of coal in India in the three years 1949 to 1951.

Production of Coal in India.

(Lakh Tons)

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
A—1949	25.7	27.1	27.0	27.0	26.0	23.8	23.8	26.3	27.3	25.9	26.0	28.8
B—1949	25.7	52.8	79.8	106.8	132.8	156.5	180.3	206.6	233.9	259.8	285.8	314.6
C—	—	—	—	—	—	—	—	—	—	—	—	314.6
A—1950	26.1	29.4	28.8	27.2	27.1	24.4	23.0	26.0	28.2	26.2	25.6	28.2
B—1950	26.1	55.5	84.3	111.5	138.6	163.0	186.0	212.0	240.2	266.4	292.0	320.2
C—	315.0	317.3	319.1	319.3	320.4	321.1	320.3	320.0	320.9	321.2	320.8	320.2
A—1951	29.1	28.9	29.6	28.5	29.7	26.7	25.3	28.2	29.1	27.9	29.3	30.7
B—1951	29.1	58.0	87.6	116.1	145.8	172.5	197.8	226.0	255.1	283.0	312.3	343.0
C—	323.2	322.7	323.5	324.8	327.4	329.7	332.0	334.2	335.1	336.8	340.5	343.0

A stands for Production, **B** for Cumulative and **C** for Moving Annual Total.

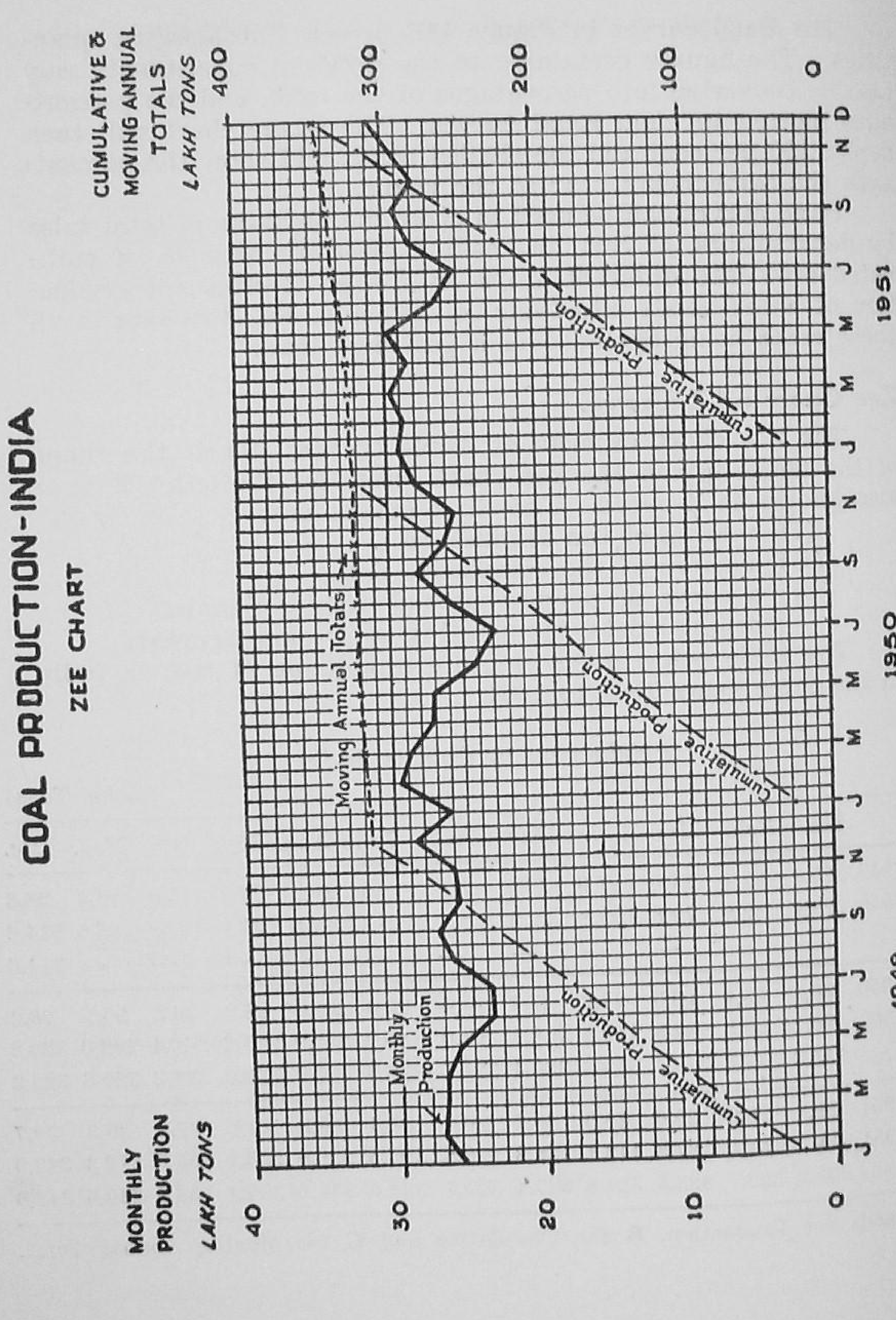

COAL PRODUCTION-INDIA
ZEE CHART
MONTHLY PRODUCTION
LAKH TONS
CUMULATIVE & MOVING ANNUAL TOTALS
LAKH TONS
40
30
20
10
0
400
300
200
100
0
Monthly Production
Moving Annual Totals
Cumulative Production
Cumulative Production
Cumulative Production
J M M J S N J M M J S N J M M J S N D
1949
1950
1951

Fig. 18C.

For each year, cumulative production figures have been calculated in the table. The moving annual total, also shown in the table, will naturally be nil for the first 11 months of 1949. The moving annual total from end of 1949 is found by dropping January production for 1949 and adding January 1950, and so on. These three sets of series are plotted in Figure 18C which gives the Zee chart showing the characteristics of coal production in India.

As will be seen from the scales denoted on the right and left of the chart, the scale used for the cumulative and moving annual total figures is ten times that of the monthly production.

In Figure 18C, fluctuations of a seasonal nature can be seen from the thick line relating to the actual monthly data. It will also be seen that the cumulative curve meets the moving annual total curve at the end of each year. The cumulative curve shows whether the total of year's production to-date is more, or less, than that of the previous year. The moving annual total curve shows the trend both month by month and year by year.

GRAPHS ON "RATIO" SCALE.

So far, we have been dealing with the natural scale graphs in which the y's are scaled in proportion to their actual values. We have seen that this method shows absolute movements in statistical series, but fails to exhibit relative movements in their proper perspective. The importance of the study of relative changes in economic investigations is growing. To study relative changes the Ratio scale or Logarithmic scale is employed as an alternative to the Natural scale.

Natural Scale contrasted with Ratio Scale.

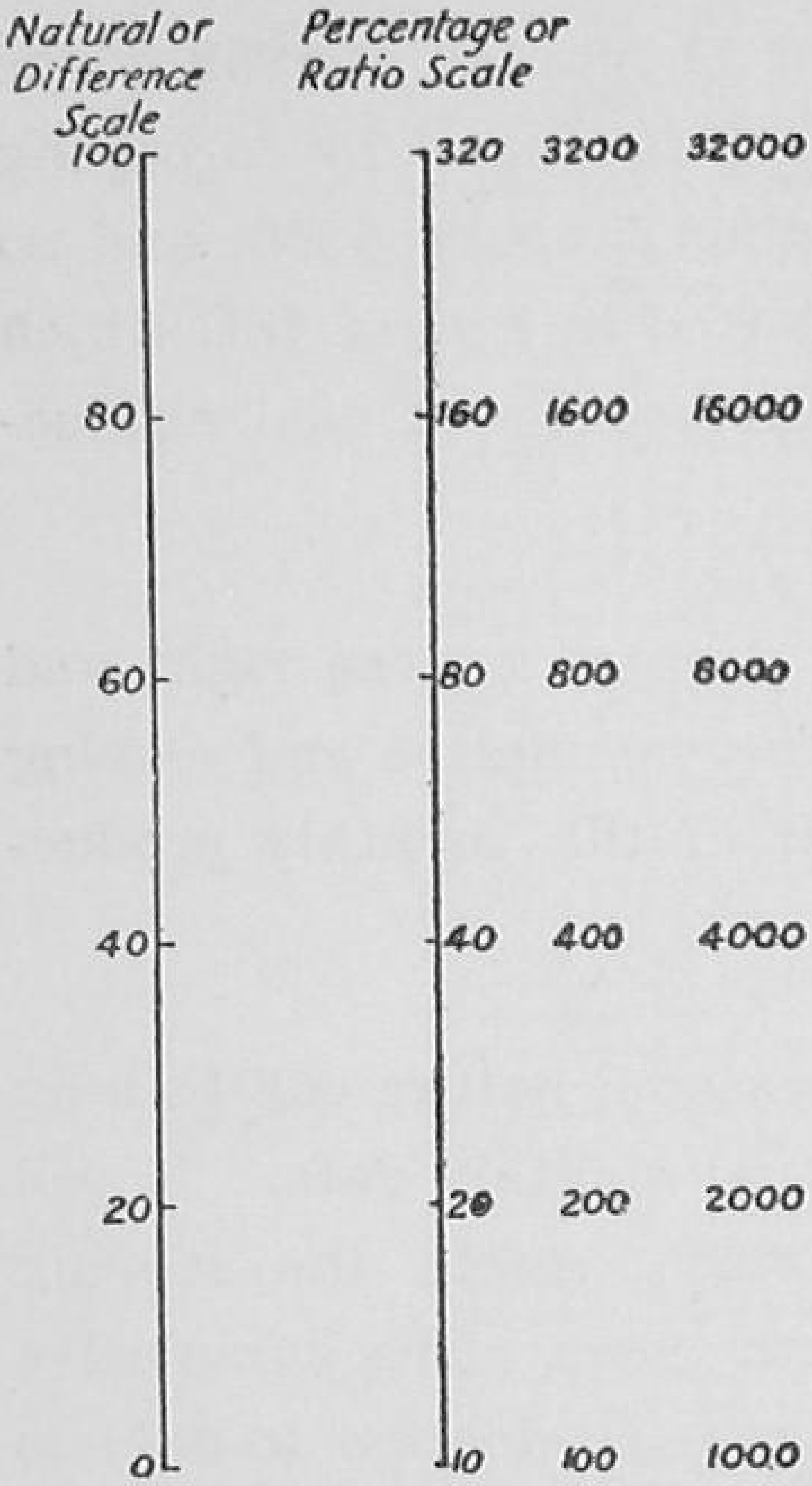


Fig. 19.

The difference between these scales is that with the natural scale equal distances on the ordinate represent equal *absolute* movements, while with the ratio scale they represent equal *proportionate* movements.

Ratio Scale.

Ratio scale is based on geometric progression, while the natural scale is based on arithmetic progression. This fact is borne out by Figure 19, which compares the natural scale with the ratio scale.

The importance and usefulness of the ratio scale can be very easily seen. Let us suppose that the population of a certain town increased as follows:

Year	Population	Increase
1920	200,000	
1930	300,000	50·0 per cent.
1940	400,000	33·3 per cent.

The absolute increase in the two periods under study is identical, that is, 100,000 in each case, but the proportional increase differs, for in the first case it is 50% of 1920, and in the latter, $33\frac{1}{3}$% of 1930. If, then, these population figures are plotted on the natural scale, the increment in population (100,000) in each case will be shown by equal distances on the vertical scale thereby leading to the conclusion that the increments between

1920 and 1930 and between 1930 and 1940 were equal. But the ratio graph will indicate that the increase took place at the rate of 50% and 33⅓% respectively.

Logarithmic Curves.

Rates of changes may be graphically presented in either of two ways:

(1) by plotting the logarithms of the given quantities on a natural scale;

(2) by plotting the given quantities themselves on a logarithmic scale.

In the former method, logarithms of the variable to be plotted on y-axis of the natural scale graph are found out, plotted, and joined by straight lines to result in the required logarithmic curve.

In the latter method, logarithms of the y variable are not found. Instead the actual values of the variable are plotted on logarithmic paper which is ruled on the ratio scale principle, illustrated in Figure 19. This method is more simple and easy than the former and has the additional merit of comparing actual values. But, if logarithmic paper is not available, the former method has to be preferred.

Table 46. *Increment of sums of Rs. 10 and Rs. 100 at 20% p. a. compound interest.*

Year	A	B
	Rs.	Rs.
1934	10	100
35	12	120
36	14·40	144·0
37	17·28	172·8
38	20·74	207·4
39	24·89	248·9
40	29·85	298·5
41	35·82	358·2
42	42·16	421·6

Logarithmic curve is also termed as *Semi-logarithmic curve* for the simple reason that one variable, y, is plotted on the logarithmic scale, while the other variable, x, remains on the natural scale.

Table 46 shows how a sum of Rs. 10 borrowed by A and another sum of Rs. 100 borrowed by B in 1934 increased at the rate of 20% per annum, compound interest being charged on both the sums.

Figure 20 shows a graph of the two series on the *natural* scale. From the figure it appears that the rate of increase in series B is more rapid than that in series A. This, however, is not the case, as is shown in Figure 21 in which the two series

Sums of Rs. 10 and Rs. 100 rising at compound interest on the Natural Scale.

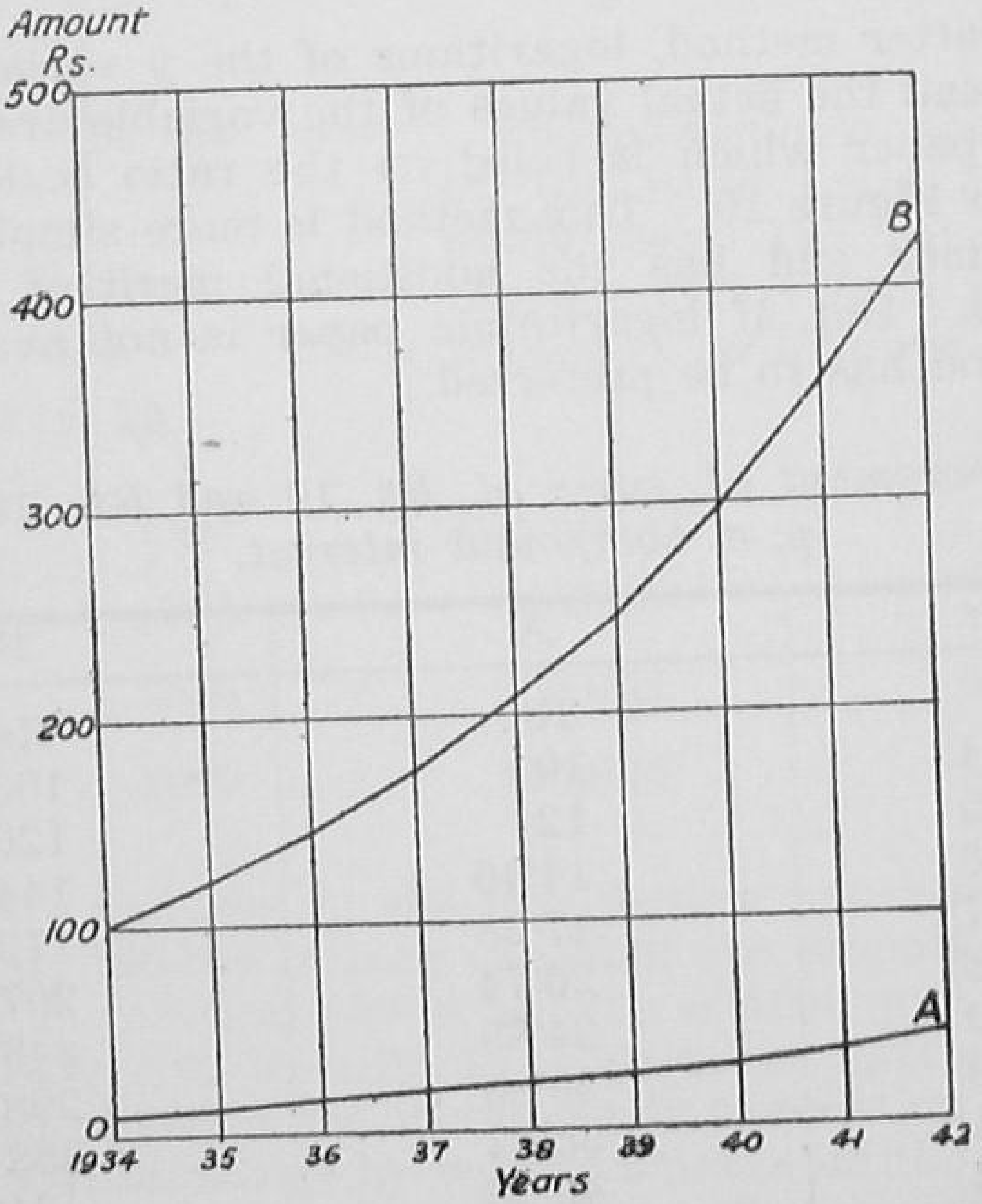


Fig. 20

are drawn on the *ratio* chart. The curves A and B in the Figure run parallel to each other, and this clearly brings out the fact that the rate of increase in both the series is identical.

It will be seen that the curves in Figure 21 have been drawn according to the first of the two methods discussed above, *viz.*, by plotting the logarithms of the two variables against time on the natural scale graph-paper. Instead, if logarithmic paper had been used and actual items were plotted on it, the resulting curves, again, would have run parallel to each other indicating a uniform rate of increase.

Sums of Rs. 10 and Rs. 100 rising at compound interest on the Ratio Scale.

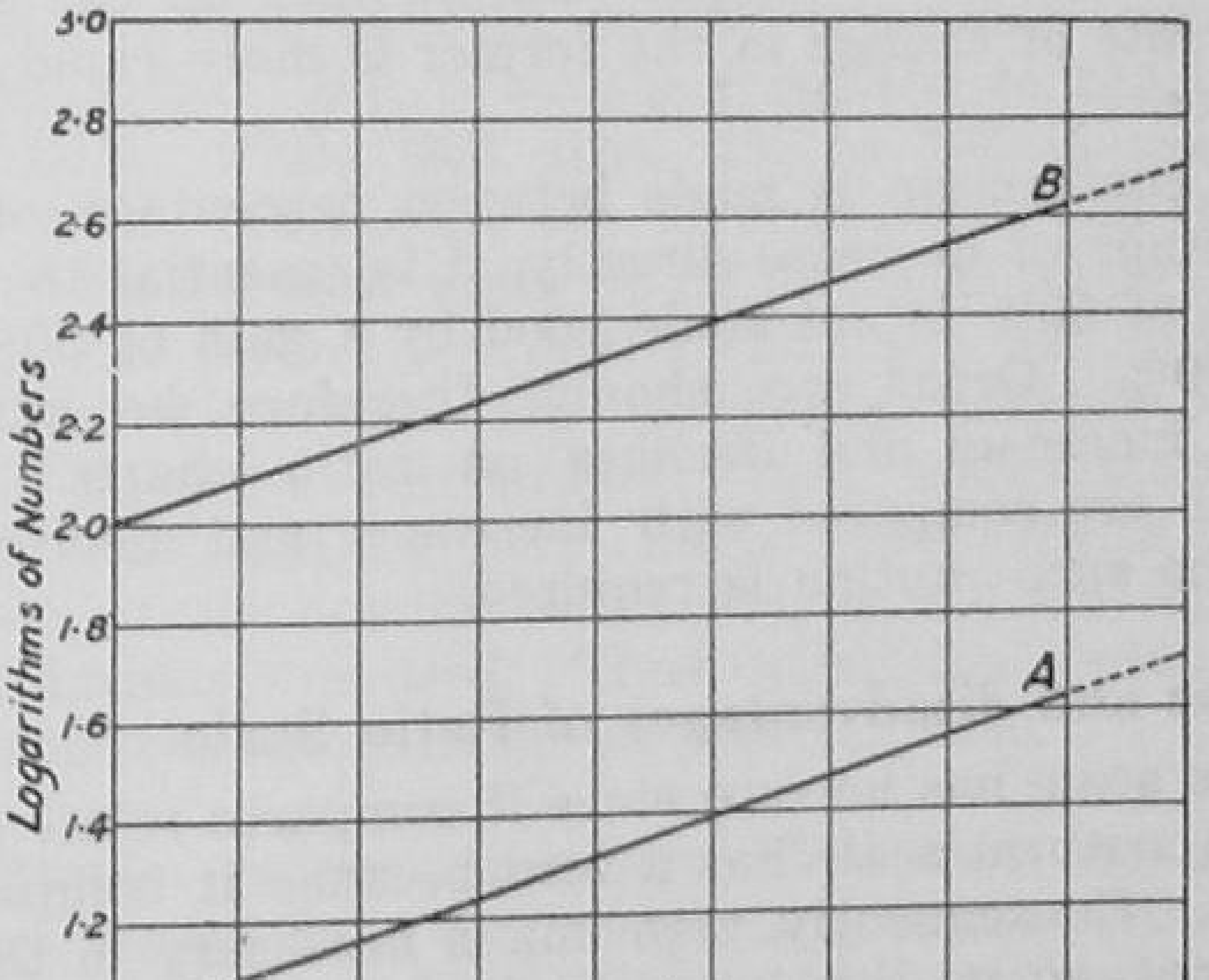


Fig. 21

Instructions for reading of Logarithmic Curves.

The following general rules will help in the reading of logarithmic historigrams:—

1. If a curve rises upwards, the rate of growth is increasing. (*cf.*, Figure 21).

2. If a curve is falling downwards, the rate is decreasing.

3. If a curve is ascending but is *nearly* straight, the magnitude it represents is growing at a nearly uniform rate.

4. If a curve is descending but is *nearly* straight, the magnitude it represents is decreasing at a nearly uniform rate.

5. If a curve is a straight line, the rate of change is uniform or constant. (*cf.*, Figure 21).

6. If a curve is steeper in one portion than in another portion, the rate of change in the former is more rapid than that in the latter.

7. If two curves on the same ratio chart are found running parallel they represent equal percentage rates of change (*cf.*, Figure 21).

8. If one curve is steeper than another on the same ratio chart, the rate of change in the former is more rapid than that in the latter.

When comparison is made between percentage of increase and percentage of decrease directly, it is essential to remember that a loss of 50% is not made good by a gain of 50%, but by that of 100%. Great care should, therefore, be exercised in comparing increases and declines on ratio charts. However, if increases are compared with increases and decreases with decreases no such caution is required.

Advantages and disadvantages of Ratio Scale.

A ratio scale has no zero since it compares relative rates of change. A natural scale has a zero because it compares absolute values. Consequently, zero line is necessary in the natural scale but not so in the logarithmic scale. Since there is no zero in a ratio scale, there is no danger of fallacious conclusions being drawn from the graph. We have seen while dealing with False Base line how fallacious conclusions might be drawn from graphs drawn on natural scale whose zero line is omitted.

Ratio scale makes **extrapolation**—finding out a future probable figure—possible, if the data are organic in character. If population figures of a certain country are given and are plotted on the ratio scale, the curve may be extended in continuation with its trend beyond the last date to the next date to obtain thereby a fairly accurate estimate of what the next

figure is likely to be. In Figure 21, the curves A and B have been extended beyond 1942 in continuation of their trends and it is possible to read from the dotted lines that the amounts at compound interest in 1943 will be Rs. 50·64 and Rs. 506·4 respectively.

The logarithmic scale is specially important in the case of **index historigrams.** They should generally be drawn on ratio scales, because index numbers are more concerned with proportionate changes than with actual ones. Index numbers plotted on the natural scale convey false impression. For example, if price index numbers for three successive years are 100, 130 and 160, each succeeding number differs from the preceding one by 30 points. This difference would be represented by equal distances on the natural scale, so that the rise in prices would appear to be equal. But, a change from 100 to 130 implies an increase of 30%, while that from 130 to 160 means a rise of 23-1|13 per cent. Therefore, the difference between the indices when plotted on the ratio scale would be shown as 30% between the first and the second year and 23-1|13 per cent. between the second and the third year. The percentage changes of the two periods will then be comparable. Such comparison, it was pointed out while dealing with index historigrams drawn on natural scale, is not possible when simple, and not logarithmic, index historigrams are used. Thus, logarithmic graph is very useful for relative comparisons in point of time.

Logarithmic graphs, however, have two disadvantages. Firstly, they are no good for comparison of the absolute sizes of different variables. Secondly, negative values cannot be shown on the logarithmic scale. It may also be added as a third disadvantage that the ordinary reader is unfamiliar with logarithms and logarithmic graphs and is, therefore, unable to interpret what ratio charts really imply.

General Remarks.

We have studied the graphical methods by which time series can be presented and changes in a single variable or in two or more variables can be studied. We have taken no account of the study and comparison of short-time and long-time

changes in a time series, which would be taken up in the next Chapter.

We have also discussed that the changes in a time series can be graphically studied in their absolute values through absolute historigrams, in their values relative to the base by index historigrams, and in their proportionate values by historigrams—absolute or index—drawn on the ratio scale. We have also studied the Mixed Graph, the Zone and Band curves and the Zee chart. We have, thus, studied the methods by which comparisons can be made in point of time. It is now proposed to deal with the graphical methods by which statistical nature of frequency distributions can be studied.

FREQUENCY GRAPHS

Frequency distribution may be discrete or continuous. A table giving frequency distribution of a group presents the data in compressed form, but many people can normally appreciate the relative sizes of a number of quantities more readily when they are graphically presented than they can do by looking at a table. Graphical presentation may, therefore, be very well employed as an addition to the method of tabulation in bringing out the statistical nature of a group, whether discrete or continuous.

Statistical Nature of a Group.

In Chapter IX are given several tables relating to frequency distribution of groups. In all of them one characteristic would be observed: It is that frequencies rise up to a certain maximum point, and begin to fall after this point is reached. Another point to be noticed is that this rise and fall shows certain regularity. The question that naturally arises is whether this regular rise and fall noted in the tables in Chapter IX is simply an arbitrary assumption, a characteristic of the particular frequency distributions referred to, or a feature common to many varieties of phenomena. It would be found that this feature is common to many or most phenomena.

Let us pick up a good number of leaves from any tree *at random,* measure their lengths, and tabulate the lengths in

certain well-defined groups. Or, let us take a few rupee coins, toss them, and see how many times only one coin falls with head upwards, how many times two fall in the same manner, how many times three, and so on. In both these cases it will be found that the frequencies begin with small magnitude, rise up to a certain maximum and begin to fall. The former is a case of natural phenomena, and the latter that of pure chance, and yet the rise and fall of frequencies would occur in identical manner in both of them.

Let us take another example. Of the 111 students admitted to the B.A. class of a certain college in a particular year, 55 students were picked up at random, their heights measured, and tabulated as in table 47.

Table 47. *Height of 55 boys arranged in ascending order.*

Serial No.	Height in inches	Serial No.	Height in inches	Serial No.	Height in inches	Serial No.	Height in inches	Serial No.	Height in inches
1	59	12	63·5	23	66	34	67	45	68·5
2	60	13	64·5	24	66	35	67	46	68·5
3	61	14	64·5	25	66	36	67·5	47	69
4	61·5	15	64·5	26	66	37	67·5	48	69
5	62	16	65	27	66·5	38	67·5	49	69
6	62	17	65	28	66·5	39	67·5	50	69·5
7	62·5	18	65	29	66·5	40	68	51	69·5
8	62·5	19	65·5	30	66·5	41	68	52	70
9	63·5	20	65·5	31	67	42	68	53	70
10	63·5	21	65·5	32	67	43	68	54	71
11	63·5	22	65·5	33	67	44	68·5	55	72

From the table it will be found that (1) the height of 67 inches is repeated the largest number of times, (2) as we proceed on either side of 67 the number of times each height is repeated is not only less than the number of times 67 inches is repeated, but also the number of times each height occurs goes on falling, so that (3) a large majority of heights group round 67 inches. 67 inches is the modal height.

From these examples we can deduce the following law which would apply to most other cases: **Phenomena tend to fluctuate about a norm known as the mode, and a large majority of items cluster round it. As the distance from the mode widens, the items become fewer.**

We can arrive at this conclusion graphically as well. We may plot the array of lengths of leaves or of the results of tossing coins. We have plotted in Figure 22 the array of the height of 55 boys. In this graph we see that near the extremes the heights change rapidly, while the fluctuations are not so marked in the middle. We further see that the mode is 67 inches since the largest number of lines stand for this number in the figure.

Array of height of boys of 17 years chosen at random.

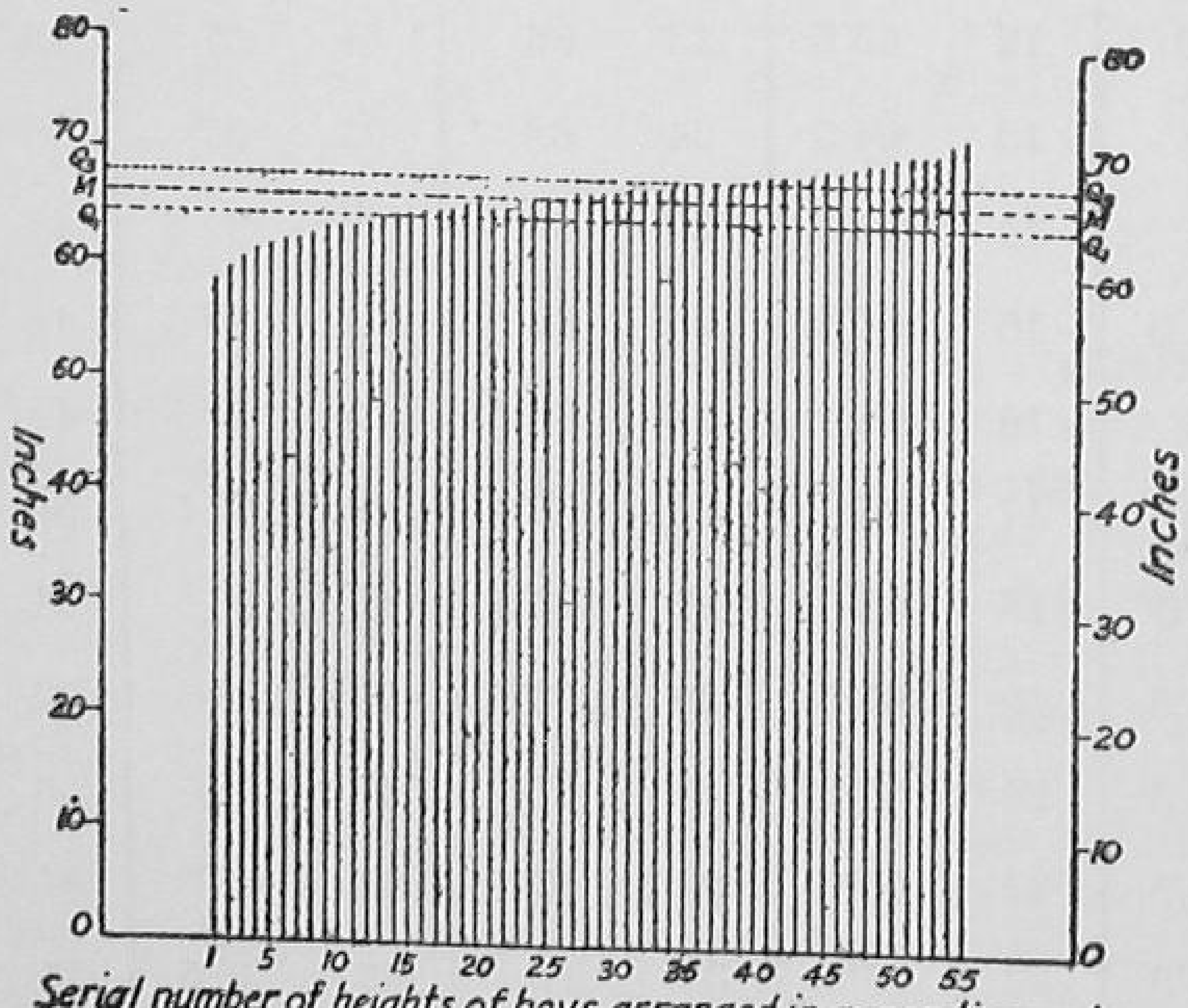


Fig. 22

One might like to know whether the same result would follow if the heights of the remaining 56 students were also measured. That is, would the location of the mode be affected by the inclusion of more items? The answer is that if the 55 boys who were selected represented a fair sample of the whole class, then the use of a larger number would give a greater regularity to the variation of the sizes, and results not materially different from the one we have arrived at would follow.

From Figure 22 it is very easy to locate the median and the two quartiles. The line drawn parallel to the base from the height of the 28th boy cuts the scale at 65.5 inches which is the required median. Similarly, quartiles can be located, as shown in the figure.

Frequency Graphs for Discrete Series.

The simplest mode of illustrating a discrete series is the **line** or **bar frequency diagram.** Size of item is taken on the hori-

Line Frequency Diagram.

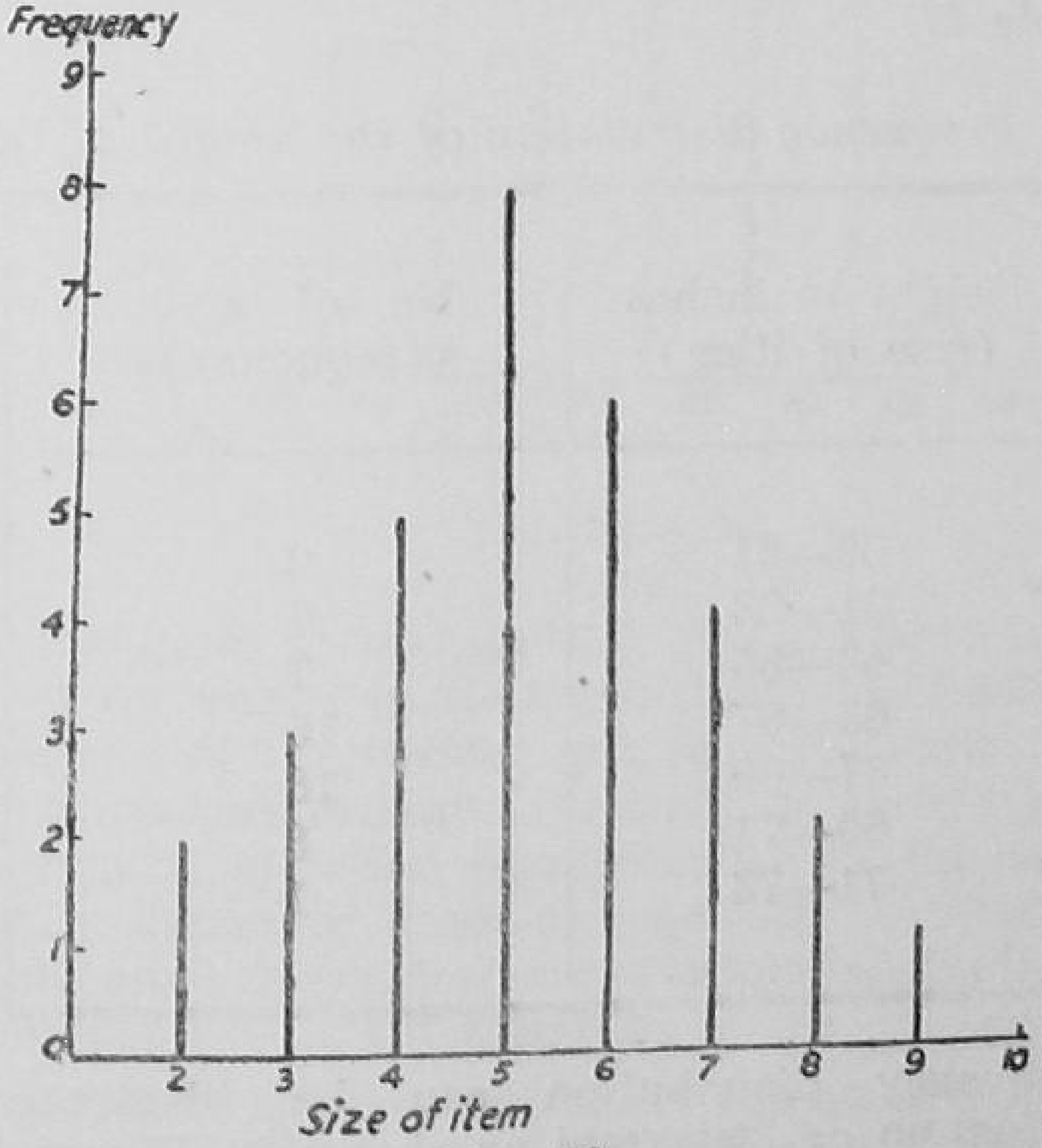


Fig. 23

F. 22.

zontal line and the frequency on the vertical scale. Table 21, page 169, gives a discrete series which is graphically presented in Figure 23. Instead of drawing simple lines as done in the Figure, bars of uniform thickness could also have been drawn to improve the appearance of the Figure. The resulting Figure would have been called *Bar Frequency Diagram.* Sizes of items should be, as they are in the Figure, separated on the horizontal scale by sufficient blank space so that the neighbouring lines may become absolutely distinct from each other.

Frequency Graphs for Continuous Series: Histogram.

If in a series the range, that is, the difference between the largest and the smallest items, is very large and instances occur at a great number of points between the two extremes, the above method of the line or bar frequency diagram is not suitable to follow, for it is impracticable to place a line at each measurement. In such cases the data must be divided into classes and each class treated as a whole. Table 48 gives such a classification of the data relating to the height of 55 boys given in Table 47.

Table 48. *Frequency distribution of the height of 55 boys.*

Height in inches (size of item)	No. of boys (frequency)
59—61	2
61—63	6
63—65	7
65—67	15
67—69	16
69—71	8
71—73	1

This frequency distribution can be illustrated by the **rectangular diagram** or **histogram** as shown in Figure 24. The histogram is composed of a set of rectangles one over each

class-interval on the horizontal scale. The heights of the rectangles are in proportion to the frequencies in the class. The area enclosed by the histogram is bounded by the lines of the ordinates, the base line and the parallels to base line at the top of each class-interval.

Histogram representing the heights of 55 *boys*

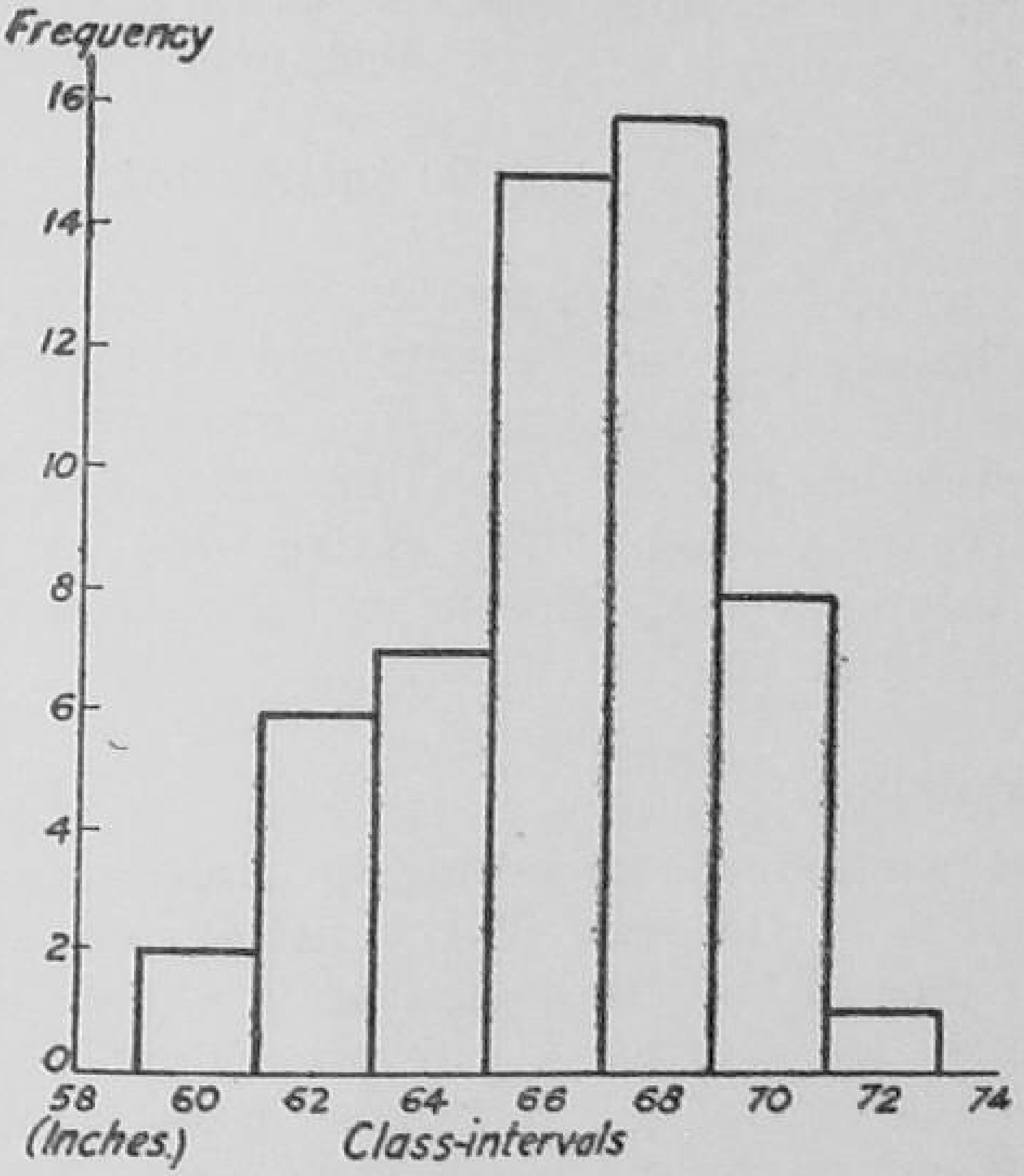


Fig. 24.

The rectangular histogram has a few characteristic features. The series of rectangles in the figure illustrates fairly accurately the relative size of the various groups. The entire distribution of frequencies among the several classes becomes at once visible. The histogram is a better representative of the height of 55 boys actually measured than any smooth curve (Figure 25) would be, although the smooth curve would be a better representative of the heights of 111 students.

The total area of the rectangular erected on each class-interval is exactly equal to the number of frequencies in that class, the unit of area being measured by a rectangle one frequency unit

in height and one class-interval in width. Thus the area of the figure equals the total number of frequencies. These two are, no doubt, significant facts of the rectangular histogram, but it is not without its defects.

One defect of histogram is that different groupings would give different shapes. If the class-intervals in Table 48 were made narrower, the steps in the histogram would decrease in size. Secondly, it suggests, for instance, that there are 2 students each $\frac{61 + 59}{2}$, i.e., 60 inches high, 6 students 62 inches high, and so on. As a matter of fact, each group consists of boys having different heights and therefore the rectangular presentation is misleading. To do away with these defects, a system of smoothing the histogram has to be devised so that a curve as typical of the entire data as possible may result. For this purpose, Frequency Polygon or Frequency Curve may substitute the histogram.

Frequency Polygon.

A simple method of smoothing is simply to connect the other extremities of the base of the histogram with the mid-points of the tops of the rectangles, as is shown by the dotted lines in Figure 25. In the Figure, the lines connecting the mid-points of the tops of the rectangles have been extended to the base at points 58 and 74 inches, the mid-points of the two rectangles outside the histogram at which the observed frequencies are zero. This procedure gives an area representation of the frequency distribution which is exactly equal to the area of the histogram. The triangular strips of area which are excluded from the histogram are equal to those formerly outside the histogram but now included in the polygon. (Compare a and $\acute{a}$, b and b', c and $\acute{c}$, etc.). Thus the area of the polygon is equal to the area of the histogram. But the area of each rectangle of the histogram is not equal to the area of the corresponding section of the polygon. For, the area cut off from the class-interval 61-63, say, has been added to the preceding class, 59-61. To this extent, the polygon may be said to have re-distributed the frequency distribution.

Frequency Polygon and Frequency Curve representing the heights of 55 boys.

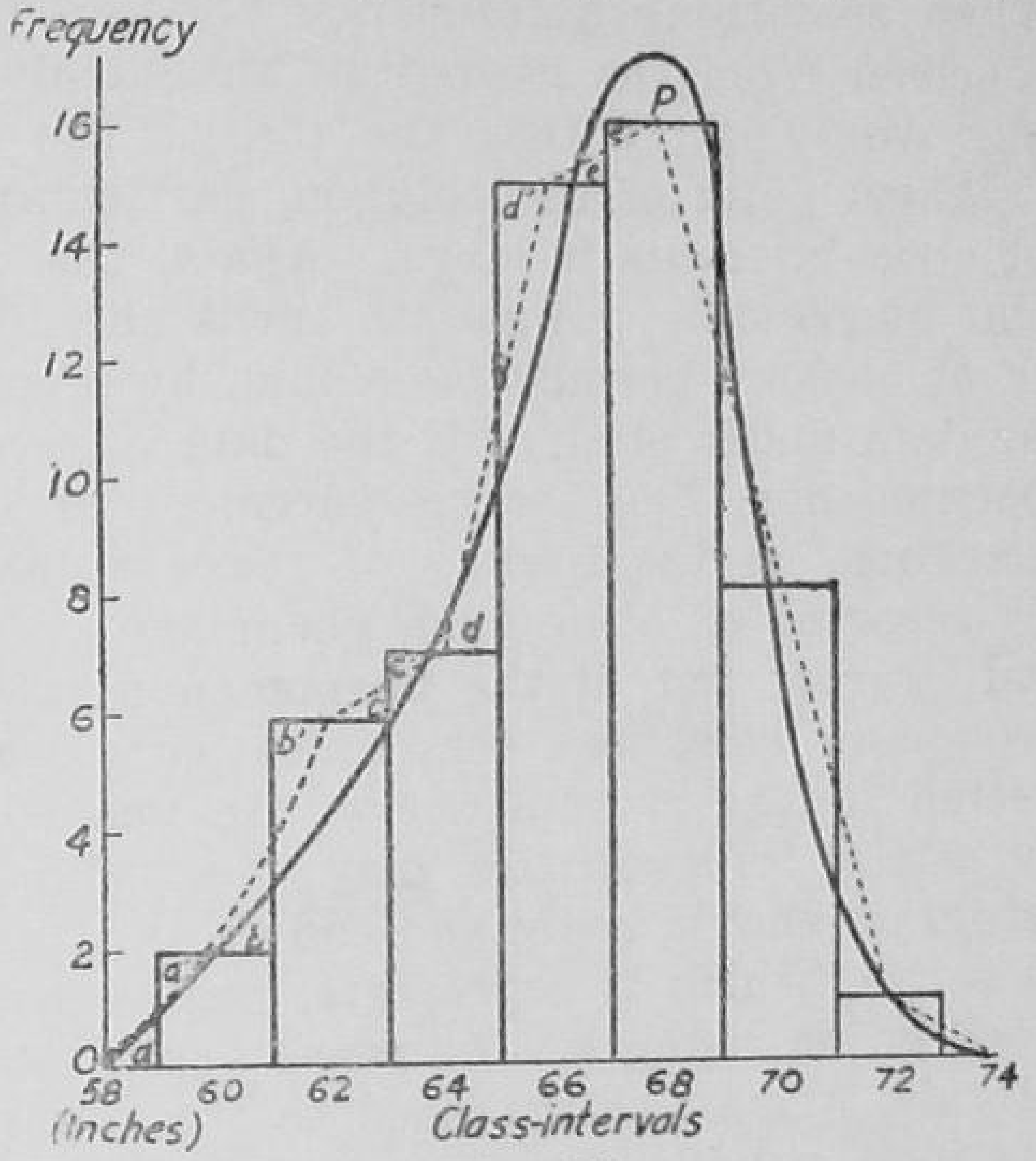


Fig. 25.

The main purpose of the polygon is to find the mode of the given series. Mode can be ascertained fairly accurately by the apex of the polygon. Apex would in all probability occur in the class-interval containing the mode, and will not be shifted greatly even if some more items are added to the series so that the original frequency distribution is modified.

The frequency polygon, however, has a defect. It shows sudden changes in the direction of the curve, particularly at the apex, P. It, therefore, fails to show regular and uniform variation in magnitude, which purpose is well served by a continuous smooth curve, called Frequency Curve.

Frequency Curve.

Frequency polygon gives the first approximation to making a continuous smooth curve. In Figure 25, frequency curve, also called **smoothed histogram**, has been drawn freehand and is shown

by the continuous line. Smoothing freehand requires great care, which, if not taken, would lead to fallacious presentation of facts. When smoothing a frequency polygon the fact that it is really derived from the histogram should always be kept in view. This would imply that the top of the curve would overtop the highest point of the polygon, particularly when the magnitude of class-intervals is large. Again, the curve should look as regular as possible; all sudden turns should be avoided. The extent of smoothing permissible would, however, depend on the particular data under study. If the data consist of records of natural phenomenon, like the measurement of leaves, or of chance phenomenon, as the tossing of rupee coins, smoothing may be freely resorted to, since such phenomena normally have a symmetrical curve; but, if the phenomenon under study is social or economic, skewness, sometimes considerably large, may be expected in the normal curve. In smoothing such a polygon only minor irregularities may be eliminated. The smoothed histogram should begin and end on the base line, since a continuous series, which it represents, begins with very few instances which go on rising but decrease again slowly to zero. As a general rule, it may be extended to the mid-point of the class-intervals just outside the histogram. Another fact that must be kept in view while smoothing a frequency polygon is that the area under the curve should represent the total number of frequencies in the entire distribution. In the matter of smoothing, experience is the best teacher.

To draw the frequency curve it is necessary first to draw the polygon. The polygon is later smoothed out. Frequency polygon may be drawn, even without first drawing the histogram, by plotting the frequencies at the mid-points of the class-intervals and joining them by straight lines. This is no doubt easy but presents difficulties in smoothing the polygon properly. Therefore, it is always safe to proceed in a sequence—first draw the histogram, then the polygon and lastly smooth it keeping in view the fact that the area of the curve should equal that of the histogram.

Frequency Distributions & Shapes of Frequency Curves.

Broadly speaking, frequency distributions are of four types. Therefore, frequency curves may take four shapes.

1. **Symmetrical Distribution**—In this type of distribution, the class-frequencies decrease to zero symmetrically on either side of a central maximum. The mean, median and mode coincide. An ideally symmetrical distribution is rare and exceptional. It occurs in biometric or anthropometric measurements.

2. **Moderately Asymmetrical Distribution**—In this type of distribution, class-frequencies decrease with markedly greater rapidity on one side of the central maximum than on the other. The curve is called "skew'.' When there is greater length of tail to the right of the maximum, skewness is said to be positive. (Fig. 2 on page 190 refers). A curve, on the other hand, is said to be negatively skewed when the tail is longer to the left of the maximum. Moderately asymmetrical curve is perhaps the most common of all smooth forms of frequency distributions. At any rate, curves relating to economic and social phenomena are almost always skew in a greater or less degree.

3. **"J-shaped Distribution**—This is an extremely asymmetrical distribution, the class-frequencies running up to a maximum at one end. This form of distribution is unusual. It may be found to occur in the distribution of wealth in the population.

4. **"U-shaped" and "V-shaped" Distributions**—In these distributions, maximum frequencies are found at either end of the range with minimum frequencies at the centre. These forms of distribution are also rare. They may be found in certain types of meteorological data.

Normal Distribution and Normal Curve of Error.

One of the recognised facts of statistical theory is that frequency distributions, particularly those relating to natural and chance phenomena, *tend* to the normal form as the size of the sample is increased; their curves are bell-shaped. According to Fisher: "A variate is said to be normally distributed when it takes all values from $-\infty$ to $+\infty$, with frequencies given by a definite mathematical law, namely, that the logarithm of the frequency at any distance d from the centre of the distribution is less than the logarithm of the frequency at the centre by a quantity proportional to d^2. The distribution is, therefore, symmetrical, with the greatest frequency at the centre; although the variation is unlimited, the frequency falls off to exceedingly small values at any considerable distance from the centre, since

a large negative logarithm corresponds to a very small number".*

An ideal symmetrical distribution may depart from the normal distribution inasmuch as the curve of the former may be flat-topped or may fall off to zero. The curve of normal distribution (also called the Normal Curve of Error and the Normal Frequency Curve) is not flat-topped; it stretches infinitely on both sides of the central maximum, closely approaching the x-axis but never meeting it.

Normal curve indicates what is expected of the phenomenon to which the curve relates, under normal conditions. The curve eliminates accidental variations and establishes normal tendencies. If such a curve has been once obtained with adequately representative sample, it can be utilised to speak for the whole universe. If more measurements are taken from the same universe, not only will they fall within the curve, but most of them will be found to cluster round the mode. Or, if the groups are re-arranged, those groups which are nearer the modal group will contain, as a rule, more cases than the groups more distant.

In coin tossing or dice throwing, one is concerned with similar and independent events and the chance of success or failure, in each individual case, is $\frac{1}{2}$ and $\frac{1}{2}$. Where these assumptions are true, the normal curve may be used to represent a frequency distribution. These assumptions are not true of economic data. In wage distribution, for instance, the wage-earners are not similar and the employment of one wage-earner affects that of others. The chance of employment or otherwise of a particular worker on a particular wage is also not $\frac{1}{2}$ and $\frac{1}{2}$. Frequency distribution of wages will thus be skew, and this skewness will not be removed by enlarging the sample of wage-earners. (Also see Chapter XX on Probability and Sampling).

Ogive Curve.

Of the three methods of presenting frequency distribution —the histogram, the frequency polygon, and the frequency curve—the last is the best for many purposes. But these methods are based on the frequencies of the class-intervals and not on the cumulative frequencies. Ogive curve is based on

*R. A. Fisher: *Statistical Methods for Research Workers*, 10th ed., p. 43.

cumulative frequencies and is, therefore, also designed as **cumulative frequency curve.**

Table 49 gives the frequency distribution of 204 weekly market quotations of the shares of the Imperial Bank of India taken week by week from 1st January 1933 to 22nd December 1936.

Table 49. *Cumulative Frequency Table showing market values of the shares of the Imperial Bank of India (Paid-up Value Rs.* 500).

Market value of shares	No. of quotations (frequency)	Cumulative frequency
1150—1200	11	11
1200—1250	44	55
1250—1300	9	64
1300—1350	10	74
1350—1400	7	81
1400—1450	6	87
1450—1500	12	99
1500—1550	30	129
1550—1600	51	180
1600—1650	20	200
1650—1700	4	204

To construct cumulative frequency curve from the table, the horizontal and the vertical scales are chosen in the same manner as in the case of histogram, polygon or curve; but the essential difference betweeen the plotting of frequency polygon and of ogive curve is that in the polygon the frequency must be plotted at the mid-point of the class-interval, while in the ogive it must always be plotted at the upper limit of the class-interval. Thus, in Figure 26 we mark 11 against Rs. 1,200, 55 against Rs. 1,250, and so on. The successive points are later connected by straight lines with a ruler. The resulting curve is an ogive curve. This curve can also be smoothed, much like the smoothing of frequency polygon, but it has not been done in Figure 26.

Ogives may be used for the purpose of comparing groups of statistics in which time is not a factor. Ogives, in general, are not easy for the ordinary person to interpret. Histograms are readily understood by him. Ogives are primarily drawn for determining medians, quartiles, percentiles etc. To determine the median of the data given in table 49, a line from the mark $\left(\frac{204}{2}\right)$ or 102 on the vertical scale is drawn parallel to the

Ogive Curve of The Frequency Distribution of the values of 204 shares of the Imperial Bank of India.

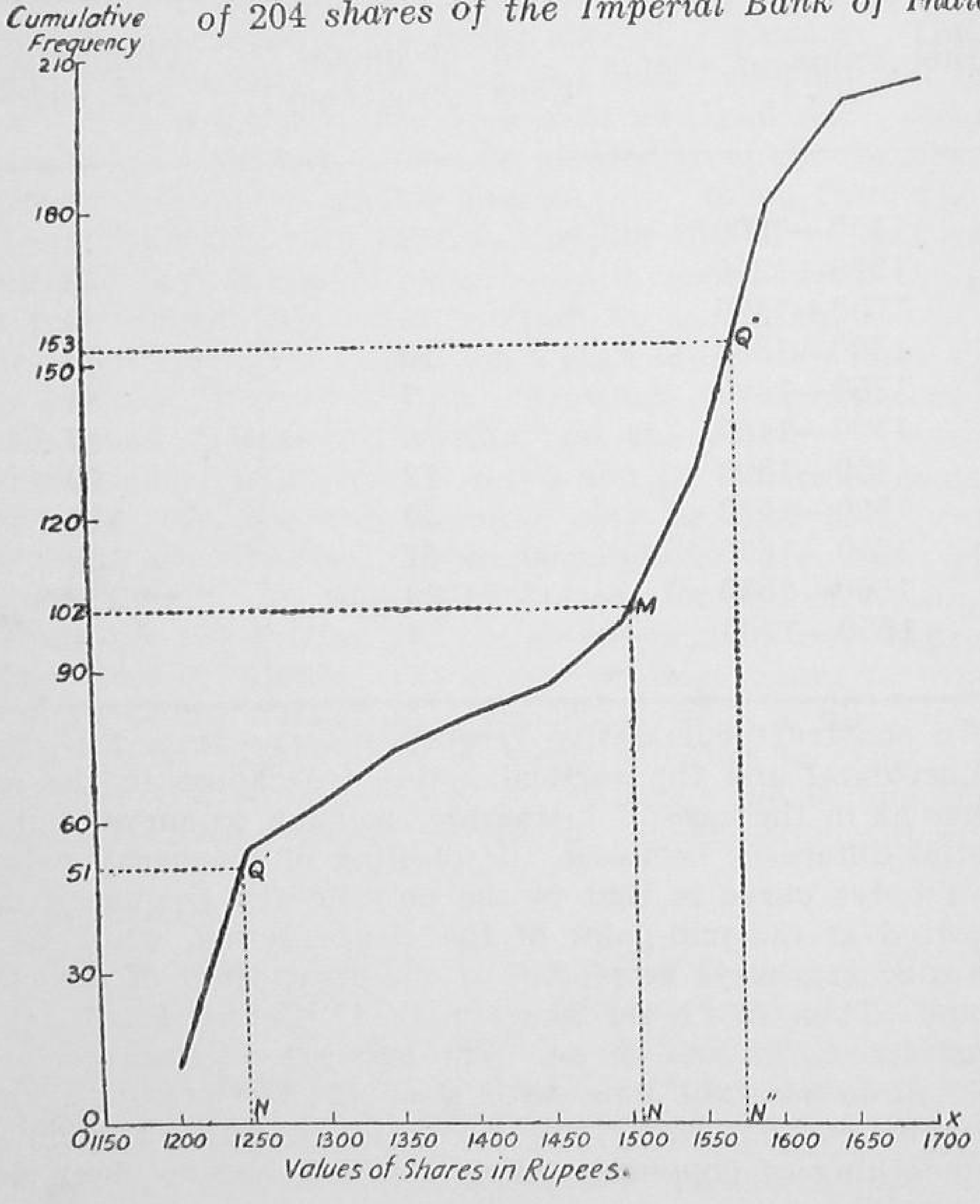


Fig. 26.

horizontal axis to intersect the ogive at M (see Figure 26), and then a perpendicular is drawn from M on OX cutting it at N. ON, read from the figure, gives the median which is Rs. 1,505. Similarly, ON″ gives the value of the upper quartile as Rs. 1,576 and ON′, the value of the lower-quartile as Rs. 1,245. Deciles and percentiles can also be likewise determined. This method of locating the median etc. is far more easy and simple than the methods discussed in Chapter IX, and is particularly so when the data given are imperfect.

The ogive is useful for yielding other results as well. Suppose an ogive represents the cumulative frequencies of income-tax payers in a certain country. We can easily find out from it the total number of tax-payers paying not less than a certain sum. Again, if data relate to wages of employees in a factory, the number of workers getting not less than a certain wage can be ascertained from the ogive. Similarly, if the government of a country wishes to formulate a scheme of graded retrenchment, this method of determining the number of employees getting not less than a certain salary would be found very useful. For, simply an ordinate need be drawn from the amount of money under consideration in the three cases to intersect the ogive, and the value of this ordinate be read on the vertical scale to know the number of tax-payers, wage-earners or government employees as the case may be. Further, the mode can also be located on the ogive as the frequencies are most numerous where the curve has the greatest tendency to run parallel to the vertical scale.

Galton's Method of Locating the Median.[1]

Galton has given a graphic method by which median can be located. The horizontal line is divided into equal parts corresponding to the unit of measurement, and vertical line is similarly divided to show the frequency. The essential point to be kept in view while plotting the curve is that every preceding measurement is made the base for the next measurement.

Table 50 reproduces the data given in table 47, in the form of frequency distribution.

[1] It should not be confused with Galton Graph, which is discussed in Chapter XVI on Regression.

Table 50. *Frequency Distribution of heights of 55 boys.*

Height	Frequency	Height	Frequency	Height	Frequency
59	1	64.5	3	68	4
60	1	65	3	68.5	3
61	1	65.5	4	69	3
61.5	1	66	4	69.5	2
62	2	66.5	4	70	2
62.5	2	67	5	71	1
63.5	4	67.5	4	72	1

Locating the Median by Galton's method in a series of heights of 55 boys.

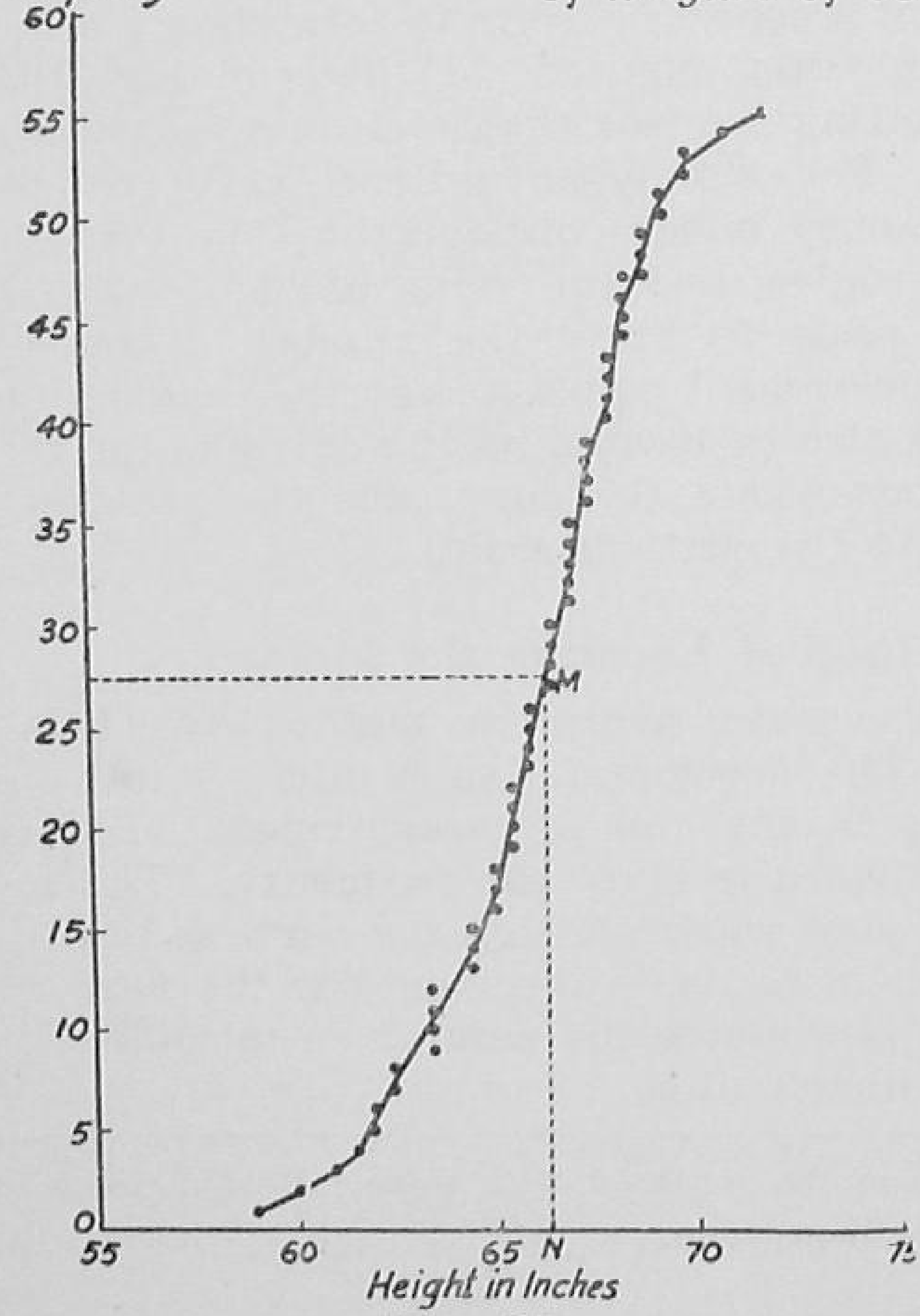


Fig. 27.

The frequency distribution given in table 50 is graphically represented by Galton's method in Figure 27. Starting with 59, one dot is put down on the ordinate through the 59th mark standing for one student. From this point a horizontal line is drawn upto 60, the next height on the ordinate through the point 60. Starting from the new base, only the dot is again marked. From this dot again a horizontal line is drawn and proceeding in the same manner as many dots are marked at each successive height as there are students of that height. Thus 55 dots are put down. When this has been done, lines are drawn to connect every two successive dots which are horizontally apart. Where the dots are odd in number, the line passes through the middle dot, while the line passes through the point midway between the two middle dots if the number of dots is even. Thus a continuous curve is obtained. To locate the median, the position of 27½th point is marked on the vertical scale, since the height of the 27½th boy is the median. From this point a horizontal line is drawn to intersect the curve. From the point of intersection a perpendicular is drawn on the horizontal base to intersect it at N. ON gives the value we desire, 66.5 inches. This is the median of the height of 55 boys.

ECONOMIC CURVES

Economic laws are very often expressed through curves. It should, however, be clearly understood that the term 'law' in economics does not have the same significance as that in exact sciences. A law in any science, exact or inexact, is no doubt a statement of general tendencies; but, unlike the exact sciences a law in economics does not state these tendencies in quantitative terms. That is, there is no single type of mathematical equation for a particular economic law. The so-called 'equation of exchange' which seeks to explain the general relationship between the quantity of money and the general level of prices does not, in fact, furnish us with any *law* connecting changes in prices with changes in the quantity of money. Other things being equal, a doubling in the quantity of money will exactly double prices. So stated, the quantity theory of money is a truism. There is no law here, because other things are hardly

ever equal, and a doubling in the quantity of money may under certain conditions produce no changes in prices, while it may more than double prices under other conditions. That is, no fixed relationship can be established between quantity of money and level of prices for all times. Likewise, no quantitative relationship of a general character can be established between changes in price and changes in demand or supply, inasmuch as it is difficult to specify exactly the equilibrium price. The equilibrium price of the same commodity may be different at different times and in different places; there may be more than one position of equilibrium for the same commodity, at the same time and at the same place. Thus, laws of demand and supply are only truism; the demand and supply curves that are usually drawn connect imaginary quantities of one variable with equally imaginary quantities of the other. The curves so drawn, no doubt, help a clearer understanding of the laws of supply and demand; they do not prove that in actual practice, the law must operate exactly in the manner shown in the curves. These curves are drawn on certain assumptions. The picture depicted by them is true only where the assumed conditions exist. A theory of equilibrium based on the assumption of free competition in a capitalistic society would not be true of State-capitalism or another form of economy. In contrast with this, the laws of exact sciences are independent of economic systems: the law of gravitation or the law of expansion of gases or metals is as true in capitalist England as in the communist U.S.S.R.

It should be clear from the above that economic laws are not universal, that they are true only when the assumptions underlying them materialise, and that, therefore, there is no particular mathematical equation governing the curves depicting these laws.

Law of Increasing Cost or Diminishing Returns.

Before drawing a cost curve, we have to assume that a certain law connects the variations in cost with those in quantity produced. In other words, we have to assume an equation that our curve would satisfy. Suppose the equation of the cost

curve we are going to draw is that of a parabola of the second order, *viz.*,

$$y=a+bx+cx^2.$$

This means that the rate of change in y for a given change in x is not equal at all points of the curve; that is, the curve is not a straight line.

Figure 27A shows a curve whose equation is

$$y=17+2x+3x^2,$$

where, x is the quantity produced and y the cost.

Now, when $x=0,\ y=17+2\times0+3\times0=17$

$x=1,\ y=17+2\times1+3\times1=22$

$x=2,\ y=17+2\times2+3\times4=33$, and so on.

Thus,

$x=$ 0	1	2	3	4	5	6
$y=$17	22	33	50	73	102	137

Measuring x along OX and y along OY, we get the curve AB which illustrates the Law of Increasing Cost or Diminishing Returns. It shows that as the quantity produced increases, the marginal cost rises. When x increases from 0 to 1, y increases by 5 units; but, when x increases from 1 to 2 (that is, again by 1 unit), y increases by 11 units; and, when x increases from 5 to 6 (that is, again by 1 unit), y increases by as much as 35 units. This relationship between quantity produced and cost underlies our assumed equation of the curve.

In Figure 27A the cost curve is a parabola of the second order. It is not necessary that all cost curves should have similar shape. A cost curve may be a straight line or a parabola of higher than the second order.

INCREASING COST OR DECREASING RETURNS

$$y = 17 + 2x + 3x^2$$

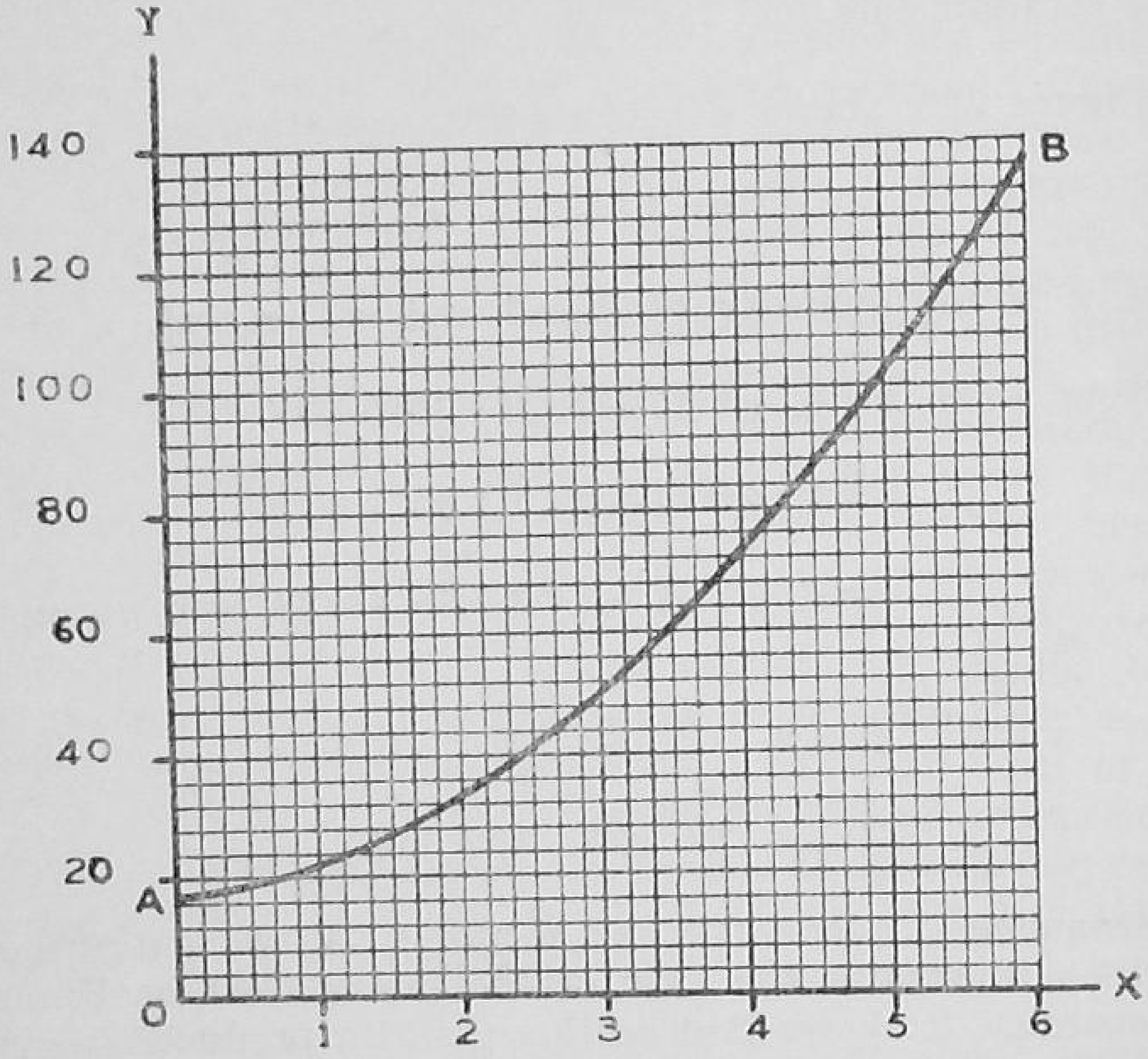


Fig. 27A.

Law of Decreasing Cost or Increasing Returns.

The curve to illustrate the Law of Decreasing Cost or Increasing Returns is logarithmic. Its equation has the form,

$$y=ab^x$$

$$\text{or, Log. } y=\log. a+x \log. b$$

The curve in Figure 27B has the equation

$$\text{Log. } y=2-0{\cdot}32x$$

$$\text{or, } y=\text{anti-log. } [2-0{\cdot}32x]$$

Now, when, $x=0$, $y=$anti-log. $2{\cdot}00=100{\cdot}00$

$x=1$, $y=$anti-log. $1{\cdot}68=\ 47{\cdot}86$

$x=2$, $y=$anti-log. $1{\cdot}36=\ 22{\cdot}91$, and so on.

Thus,

$x=$ 0	1	2	3	4	5	6
$y=$100	47.86	22.91	10.96	5.25	2.51	1.20

Demand and Utility Curves.

Measuring x along OX and y along OY, we get the curve AB which shows that y decreases with x, that is, cost goes down as the quantity produced increases.

Figure 27B also shows the shape of Demand curve and Utility curve. The values of y for given values of x calculated above, may be used to draw these curves too. In the case of Demand curve, x will show the price and y the quantity demanded at each stated price. In the case of Utility curve, AB will show a fall in utility with an increase in consumption units.

DECREASING COST OR INCREASING RETURNS

$y = \text{Anti-Log}\,[2 - 0{\cdot}32\,x]$

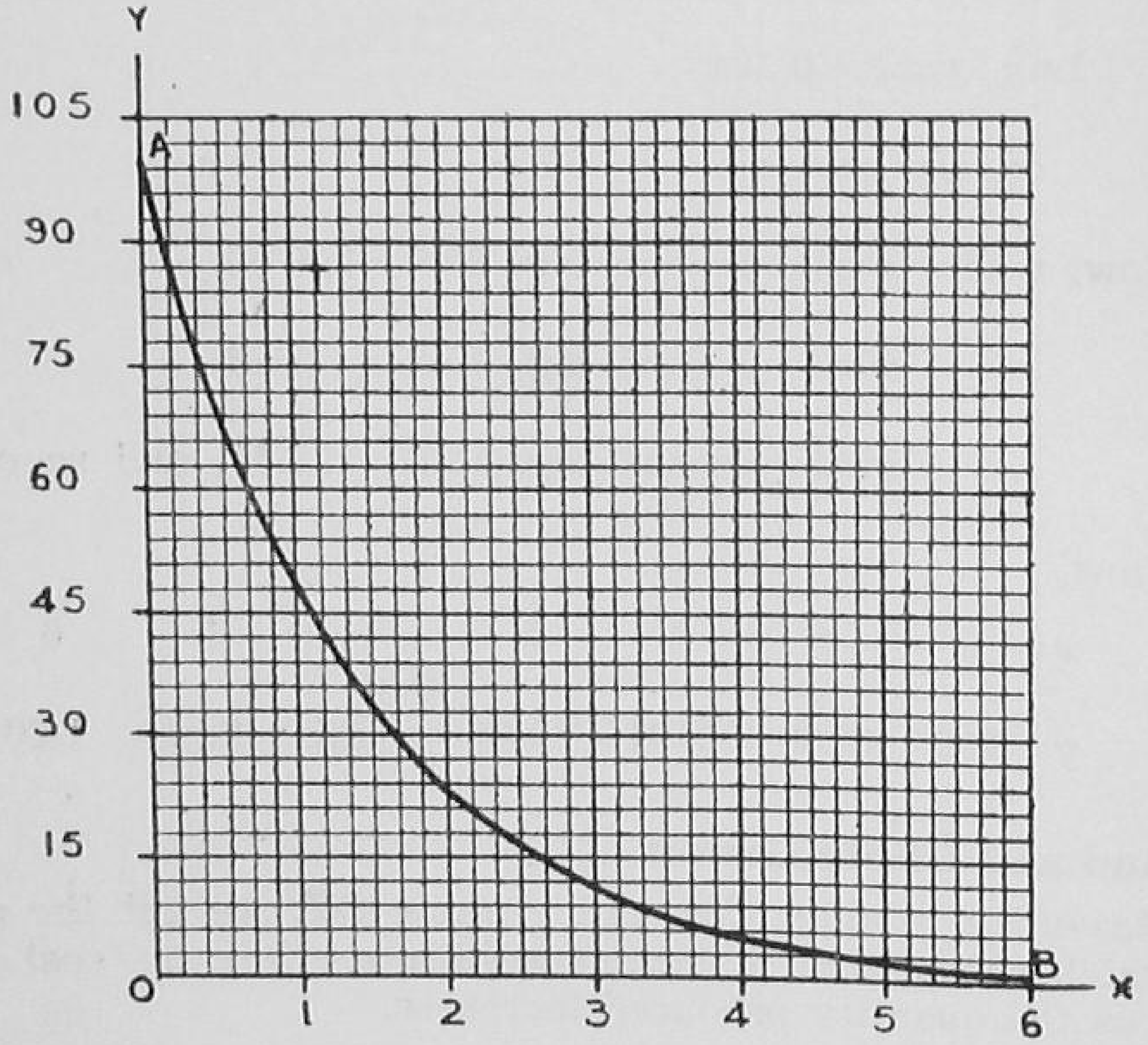


Fig. 27B.

EXERCISES

(1) 'The wandering of a line is more powerful in its effect on the mind than a tabulated statement.' Elucidate.

(2) Explain with the help of diagrams, the difference between a frequency polygon, histogram, frequency curve and ogive.

(M.A., Patna, 1942).

(3.0) Differentiate clearly between the natural scale and the logarithmic scale used in graphical presentation of data.

Draw a curve to present the Calcutta index numbers, given in exercise 15 in the Chapter On Correlation, using the logarithmic scale.

(B. Com., Alld., 1944).

(3.1) Comment on *any two* of the following statements:—

(*a*) A statistician is not an alchemist expected to produce gold from any worthless material.

(*b*) Averaging is the process of condensation.

(*c*) Phenomena tend to fluctuate about a norm with a large majority of items clustering round it.

(B. Com., Alld., 1946).

(3.2) What statistical technique and calculations are necessary to give a comprehensive idea of a distribution? Illustrate your answer.

(M.A., Alld., 1951).

(4) The following table gives the value of Imports and of Exports of (undivided) India for the years 1920—21 and 1921—22 in crores of Rupees:

Months	1920-21		1921-22	
	Imports	Exports	Imports	Exports
April ..	22	28	26	18
May ..	24	28	21	20
June ..	26	23	19	17
July ..	28	21	18	17
August ..	31	20	21	20
September ..	29	22	20	20
October ..	32	21	23	18
November ..	32	19	26	20
December ..	32	20	23	22
January ..	31	19	28	23
February ..	25	18	20	22
March ..	24	19	21	28

Plot the above figures on a graph paper, and show also the balance of trade.

(B. Com., Alld., 1938).

(5) Study the following table, and draw a graph on logarithmic scale to show net supplies of cereals and adult population in (undivided) India:—

Year	Production of Cereals	Seed and Wastage	Net Imports (+) or Exports (—)	Adult Population
	'000 tons.	'000 tons.	'000 tons.	'000
1935-36	54,177	6,722	+1,793	290,846
1936-37	59,578	7,445	+1,245	294,917
1937-38	58,739	7,342	+ 630	298,987
1938-39	54,468	6,809	+1,044	303,058
1939-40	57,244	7,156	+2,221	307,128
1940-41	54,808	6,851	+ 963	311,198
1941-42	56,550	7,069	+ 432	315,269
1942-43	58,726	7,341	— 292	319,399
1943-44	62,925	7,865	+ 298	323,410
1944-45	59,527	7,441	+ 693	327,481

(B. Com., Alld., 1946).

(6) What are the advantages of the Ratio Scale over the Natural Scale ? Plot the following data graphically on the logarithmic scale:—

Year		Total Notes issued in crores of Rs.	Notes in circulation in crores of Rs.
1933-34	..	177	167
1934-35	..	186	172
1935-36	..	196	167
1936-37	..	208	192
1937-38	..	214	185
1938-39	..	207	187
1939-40	..	252	237
1940-41	..	269	258
1941-42	..	421	410
1942-43	..	650	625

(B. Com., Nagpur, 1943).

(7) The following table shows the total sales of gold by the Bank of England on foreign account. Represent the data graphically on the logarithmic scale:—

Year			£'000
1910	..	..	14,488
1911	..	..	8,228
1912	..	..	9,670
1913	..	..	7,943
1914	..	..	8,027
1915	..	..	43,076
1916	..	..	2,360

(B. Com., Alld., 1932).

(8) Represent graphically the data given below on a single sheet of graph paper to bring out clearly the relative fluctuations in the prices of various articles. Draw such conclusions as you can from the graphs.

Wholesale prices in Cawnpore.
(in rupees per maund)

Year	Rice	Wheat	Linseed	Gur	Cotton	Tobacco
1928 ..	7.3	7.7	7.0	6.5	34.1	17.3
1929 ..	7.7	5.5	8.0	7.3	29.8	17.1
1930 ..	5.8	3.6	6.5	6.2	17.3	14.5
1931 ..	4.1	2.7	4.2	4.2	13.3	11.6
1932 ..	4.3	3.4	3.5	3.5	14.8	4.9
1933 ..	4.1	3.2	3.4	3.1	12.9	4.9
1934 ..	3.7	2.8	3.6	4.1	13.2	5.7

(M. Com., Alld., 1943).

(9) Show the results of working of Class I railways in (undivided) India graphically and comment thereon.

(In millions of £)

	Capital Outlay	Gross Earning
1923-24	464	70
1924-25	473	74
1925-26	487	73
1926-27	505	72
1927-28	594	86
1928-29	599	86
1929-30	617	84
1930-31	627	77
1931-32	631	71
1932-33	638	70
1933-34	635	72

(B. Com., Agra, 1940).

(10) Below are given the highest and the lowest prices per pound of Indian tea at London tea auctions.

Date of Auction	Highest d.	Lowest d.	Date of Auction	Highest d.	Lowest d.
20-3-52	38.65	38.51	12-6-52	35.87	34.79
27-3-52	38.51	37.75	19-6-52	36.66	35.06
3-4-52	38.39	36.32	26-6-52	34.95	34.64
10-4-52	36.81	35.97	3-7-52	34.39	33.99
24-9-52	35.24	34.88	10-7-52	34.44	32.29
1-5-52	34.62	34.03	17-7-52	31.82	29.83
8-5-52	34.54	32.84	24-7-52	31.95	29.22
15-5-52	32.22	31.77	31-7-52	30.91	29.70
22-5-52	32.92	32.80	14-8-52	33.10	29.15
29-5-52	34.19	32.73	21-8-52	31.66	30.50

Plot the data by means of Zone curve.

(11) Draw Band curves to present the following data:

Revenue Estimates of the Government of India

In Crores of Rupees.

Year	Customs	Union Excise Duties	Net Income Tax Receipts	Other Heads of Revenue	Total
1948-49	126	51	140	55	372
1949-50	125	68	115	42	350
1950-51	157	68	126	60	411
1951-52	232	86	135	62	515
1952-53	177	80	113	49	419
1953-54	170	94	105	70	439

(12) *Distribution of firms in Woollen and Worsted Industries in Yorkshire, according to number of operatives:*

Operatives	No. of firms	Operatives	No. of firms
1— 20	380	301—340	24
21— 60	320	341—380	18
61—100	182	381—420	16
101—140	147	421—460	11
141—180	92	461—500	9
181—220	66	501—700	19
221—260	39	701—900	15
261—300	30	901—	16
	Total Number of firms	..	1384

Represent this distribution graphically (by means of a cumulative diagram) and from this graph estimate the median and quartiles of the group.

(B. Com., Luck., 1930).

(13) Present the following figures relating to monthly imports (volume and value) of liquor into undivided India graphically so as to show their fluctuations, and give necessary comments.

1941-42 Months	Volume (000) Gals.	Value Rs. (000)
April	463	2648
May	395	1932
June	358	2113
July	415	2655
August	330	2104
September	363	2339
October	456	3105
November	349	2319
December	209	1526
January	230	2107
February	159	1667
March	352	1993

(14) Plot the following figures relating to population of India (undivided) so as to show the proportionate increase in population from one period to another.

Year			Population (000,000's omitted)
1872	..	..	210
1881	..	..	250
1891	..	..	290
1901	..	..	295
1911	..	..	315
1921	..	..	320
1931	..	..	350
1941	..	..	390

(B. Com., Nag., 1945).

(15) Write a short account of the use of graphic methods in statistics. Draw a diagram to represent the data given in the following table showing the number of rooms measured, in a certain investigation, in which the size lay between the limits given in the column on the left. (Area calculated to the nearest square foot):—

Area of room (sq. ft.)	Number	Area of room (sq. ft.)	Number
20 and under 40	3	200 and under 220	18
40 ,, ,, 60	14	220 ,, ,, 240	12
60 ,, ,, 80	16	240 ,, ,, 260	3
80 ,, ,, 100	36	260 ,, ,, 280	2
100 ,, ,, 120	31	280 ,, ,, 300	2
120 ,, ,, 140	35	300 ,, ,, 320	2
140 ,, ,, 160	35	320 ,, ,, 340	..
160 ,, ,, 180	15	340 ,, ,, 360	..
180 ,, ,, 200	26	360 ,, ,, 380	1

Total number of rooms measured = 251.

Read off the median and quartiles from your diagram and check your results by actual calculation.

(B. Com. Hons., Andhra, 1944).

(16) Describe the Lorenz graph. How does it differ from an Ogive? Illustrate your answer by fitting a Lorenz curve and an Ogive to the following data:—

Percentage of age distribution of the male population in British India, 1931.

Age-Group	Males
0—10	28·0
10—20	20·9
20—30	17·7
30—40	14·3
40—50	9·7
50—60	5·6
60—70	2·7
70—over	1·1
Mean age	23·2

(M.A., Patna, 1940).

(17) The following table gives the proportion of married women in 1910 and in 1920 from women of every age. Show graphically that the increase was most marked for the women of younger years.

Age	Women, Per cent. married 1910	1920
18	17.0	19.2
20	36.2	38.4
22	50.7	52.9
24	62.0	64.2
25	65.7	67.8

(B. Com., Nag., 1944).

(18) The following table gives the age distribution of widows in undivided India (Census Report 1931). Draw a graph showing the number of widows younger than any given age, and from the graph read of the median age of the widows and also the upper and the lower quartiles.

Years	No. of widows
0—10	135,862
10—20	718,101
20—30	2,456,835
30—40	4,847,631
40—50	6,480,259
50—60	5,908,159
60—70	3,743,615
70 and over	1,957,506
Total	26,247,968

(M.A., Alld., 1942).

(19) Locate the median of the following figures by Galton's Method:—

Length of Nim leaves in inches

1.35, 1.35, 1.6, 1.6, 1.7, 1.7, 1.9, 1.6, 1.5, 1.9, 2.0, 2.3, 2.6, 2.8, 2.5, 2.3, 2.9, 3.4, 3.7, 2.9, 3.2, 3.4, 2.5, 2.8, 2.8, 2.6, 2.5, 2.3, 2.4, 2.7, 2.7, 2.7, 2.9, 3.3, 1.8, 1.6, 1.5, 1.9, 1.5, 1.6, 3.4, 2.7, 3.9, 3.5, 2.9, 2.1, 2.2, 2.3.

(20) In the following table are given the quantity of white (bleached) cotton cloth imported into India (undivided) and the price per yard. Bring out, graphically, the relation between price and quantity imported year by year, and comment on the relation indicated.

Years		Cotton cloth imported million yds.	Average price per yard Rs. a. p.
1924—25	..	549	0 6 0
—26	..	465	0 5 6
—27	..	571	0 5 0
—28	..	557	0 5 0
—29	..	554	0 4 6
—30	..	474	0 4 6
—31	..	272	0 3 9
—32	..	280	0 3 0
—33	..	412	0 3 0

(21) Following table gives the production of sugar in Cuba, Java and (undivided) India during 1930-1939 in millions of quintals. Represent the figures by a suitable diagram and comment on their relationship.

Year		Cuba	Java	India
1929—30	..	44	29	17
1930—31	..	30	28	20
1931—32	..	25	26	24
1932—33	..	19	14	28
1933—34	..	22	6	30
1934—35	..	25	5	31
1935—36	..	25	6	36
1936—37	..	29	14	40
1937—38	..	29	14	32
1938—39	..	26	15	27

(M.A., Patna, 1943).

(22) The following table gives the prices of gold and wheat and net export of gold during the years 1931-32 to 1938-39:—

Years		Average price of gold (per tola) Rs. a.	Average price of wheat (per maund) Rs.	Net export of gold Crores of Rs.
1931—32	..	25 4	3.3	58
1932—33	..	30 12	3.3	65
1933—34	..	33 6	2.8	57
1934—35	..	35 8	3.1	52
1935—36	..	35 4	3.2	37
1936—37	..	36 0	3.9	28
1937—38	..	36 6	3.0	16
1938—39	..	37 12	3.4	13

Plot the above figures on a graph paper, and comment upon the relationship.

(M.A., Agra, 1943).

(23) Present the following data on graph paper and comment on the picture the resulting figure would reveal:

Area, Production and Yield Per Acre of Cotton in India.

Year		Area (million acres)	Production (million bales of 392 lbs. each)	Yield per acre (lbs.)
1947—48	..	10.66	2.19	80
1948—49	..	11.29	1.77	61
1949—50	..	12.17	2.63	85
1950—51	..	14.56	2.97	80
1951—52	..	16.21	3.13	76

(24) Draw Frequence Curves of the following distributions, and comment on their characteristics as revealed by the shapes of the curves.

Class-interval	Frequency Distribution (a)	(b)	(c)	(d)	(e)
0— 1	1	6	240	10	350
1— 2	10	41	120	125	175
2— 3	45	124	60	225	125
3— 4	120	220	30	210	75
4— 5	210	257	15	175	50
5— 6	252	205	10	130	25
6— 7	210	114	15	130	18
7— 8	120	44	30	230	7
8— 9	45	11	60	130	5
9—10	10	2	120	30	2
10—11	1	0	240	10	2

(25) Present the following data through Zee charts:

Production of Paper & Paper Boards in India.

('000 tons)

Year	Jan.	Feb.	Mar.	Apr.	May.	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
1949	8·1	7·7	8·7	9·1	9·0	9·1	9·1	9·0	9·0	8·4	8·2	8·9
1950	8·3	8·5	9·1	8·7	9·5	8·7	9·3	9·4	9·4	9·0	9·3	9·7
1951	10·0	9·9	11·0	10·5	11·2	11·0	10·9	11·1	11·3	11·4	11·4	12·1

(26) Below are given the area, production and yield per acre of tobacco in India. Draw a Mixed Graph to present the data.

Year		Area (million acres)	Production (million tons)	Yield per acre (lbs.)
1946—47	..	0.85	0.27	716
1947—48	..	0.83	0.23	634
1948—49	..	0.80	0.26	711
1949—50	..	0.86	0.26	688
1950—51	..	0.90	0.26	653
1951—52	..	0.76	0.23	660

(27) Draw an Increasing Cost curve the equation to which is

$$y=15+2x+3x^2$$

(28) (a) Draw a curve to show that y decreases with x in terms of the following equation:

$$y=100-2x-3x^2$$

(*b*) Explain the essential difference between curves showing economic laws and those showing the laws of exact sciences.

(29) Assuming that x represents the market prices of a commodity and y the quantities of the commodity demanded at the given prices, construct a Demand Curve satisfying the following equation:—

$$\text{Log } y=2-0.3x$$

$$\text{or } y=\text{anti-log } [2-0{\cdot}3x].$$

For x you may assume the values of Re. 1/-, Rs. 2/-, Rs. 3/-, Rs. 4/- and Rs. 5/- to arrive at the corresponding values of y before plotting the Demand Curve.

(M. Com., Raj., 1952).

(30) Write short notes on the following:

(1) Histogram,
(2) Historigram,
(3) Frequency Polygon,
(4) Normal Curve of Error,
(5) Ogive,
(6) Log Curve,
(7) Zee Chart,
(8) Band and Zone Curves.

CHAPTER XIV

ANALYSIS OF TIME SERIES

In examining the changes in a certain quantity over a period of time, we are concerned with two things: the interpretation of these changes, and with finding how they are related to similar changes which are observed in other time series. For instance, when we examine the series of index numbers giving the relative fluctuations of retail price of wheat in India (1873=100), given in column 2 of Table 51 we naturally ask ourselves firstly, to what the changes taking place are due and secondly, how they are related to changes in other series. We take up the first question, first.

Trend, Seasonal and Cyclical Fluctuations.

Upon perusal of the annual average indices we find that there is, on the whole, a gradual increase in the retail price of wheat and that there are sudden breaks, sometimes of large and at other times of small number of points, in this gradual increase. We know that the series of retail prices is a resultant of a large number of causes of different kinds, *e.g.,* weather conditions, transport facilities, consumers' demand for wheat, demand and prices of substitutes of wheat, and so on. Let us consider the nature of these causes with a view to determining their effects.

The first cause, and a fundamental one, is that with increase in population of India the demand for wheat, as for so many other commodities, has been growing. Therefore, there is a certain *growth factor* which is effecting a general and gradual rise in the series. The resulting gradually changing nature of the retail price of wheat is referred to as the **secular trend** of the series and this trend should be supposed to be linked up with the growth factor.

Just as the growth factor causes a rising tendency in the indices of retail price of wheat in India, the *decline factor* in some other series, say, production of indigo in India over a

series of years, would effect a falling tendency. The downward trend in such a series would then be said to be linked up with the decline factor.

Secondly, operating along with the growth or decline factor is a group of causes which do not operate continuously, but in a regular spasmodic manner. One among these causes is the *seasonal factor*. Seasons occur in the same way every year: and prices are generally low after the harvest owing to plentiful supplies, and high some time before the sowings owing to paucity of market arrivals at the fag-end of the crop season; May and June constitute marriage season in India affecting the prices of bullion and other goods. The seasonal factor causes a regular up and down movement in the series relating to the phenomenon affected by the factor. This movement is referred to as the **seasonal movement**. If weekly or monthly retail price indices were considered, wheat prices would be found to be at their lowest in March-April-May each year due to the peak marketing season, and highest in January-February. This downward and upward seasonal movement would be superimposed on the secular trend.

Another cause in this group, operating in a regular spasmodic manner, is the *cyclical factor*. During the 19th century a fairly regular up and down movement has been noticed in a good number of time series of economic data. These movements have been repeated at intervals of years ranging from 7 to 11, that is, these movements have occurred in a cycle. There are "boom" years in which the movement of an observed phenomenon may be found to be reverse of that in "depression" or "crisis" years. Retail price of wheat in our example is also affected by this "trade cycle" or **cyclical movement**.

Lastly, in addition to the group of causes producing regular up and down movement in a time series, there is another group which operates in an *irregular manner*. It includes such events as floods, raids of locusts, fires, earthquakes, wars, revolutions, and so on, which ruin the crops of the areas affected by them and raise prices. It also includes such chance

Table 51. *Index Numbers of Retail Price of Wheat, (1873=100), and Trend by the Method of Moving Averages.*

1 Year	2 Annual average Index	3 5 Yearly moving average	4 10 Yearly moving average	5 10 Yearly moving average (centred) *i.e.*, trend	6 Deviation from trend
1873	100				
4	94				
5	81	91			
6	78	100			
7	102	113			
			108		
8	147	121		108	+39
			108		
9	158	124		108	+50
			108		
1880	118	124		108	+10
			108		
1	96	115		110	—14
			111		
2	101	102		112	—11
			113		
3	103	96		112	—9
			110		
4	91	97		108	—17
			106		
5	89	101		106	—17
			106		
6	103	105		108	—5
			110		
7	121	111		113	+8
			115		
8	123	116		117	+6
			118		
9	118	123		119	—1
			119		
1890	117	129		121	—4
			122		
1	137	130		125	+12
			127		

1 Year	2 Annual average Index	3 5 Yearly moving average	4 10 Yearly moving average	5 10 Yearly moving average (centred) *i.e.*, trend	6 Deviation from trend
2	151	127		131	+20
			135		
3	127	127		136	—9
			137		
4	105	130		139	—34
			140		
5	116	141		143	—27
			145		
6	153	145		147	+6
			148		
7	206	152		148	+58
			147		
8	143	164		147	—4
			147		
9	140	165		148	—8
			149		
1900	176	152		150	+26
			151		
1	160	149		152	+8
			152		
2	142	146		150	—9
			148		
3	129	139		152	—23
			156		
4	123	138		159	—36
			162		
5	112	143		162	—20
			162		
6	155	163			
7	168	179			
8	226	184			
9	203				
1910	170				

combinations of wind, sunshine and rain in a certain season as may result in a bumper crop and, therefore, low prices. While these causes do operate from time to time, there is no regularity in their operation. Retail price of wheat is also affected by such **irregular fluctuations.**

We may conclude, then, that in analysing any given time series we look for three kinds of movement, *viz.,*

1. General (or, Secular) Trend.
2. Regular Fluctuations :
   - (i) Seasonal.
   - (ii) Cyclical.
3. Irregular Fluctuations.

Long-period and Short-period Changes.

General trend refers to secular or long-period changes. Regular and irregular fluctuations refer to periodic changes lasting a short-period of time. Therefore, changes in time series may be spoken of as (1) long-time and (2) short-time. Long-time fluctuations include secular changes ; while short-time oscillations may be classified into (i) seasonal (ii) cyclical and (iii) irregular fluctuations.

The primary task in analysing a time series is to measure and isolate long-time and short-time changes. The short-time changes may then be broken up into regular and irregular fluctuations.

Measuring and Isolating Time Changes.

In order to study any one of these changes by itself, it seems necessary to follow the method of the physicist who allows only one factor to vary at a time and eliminates all other factors. But, the statistician can rarely control the conditions of his experiment and has, therefore, to be satisfied with ridding, so far as possible, the recorded data of the apparent effects of the extraneous causes. If we desire to study the long-time changes in the retail price of wheat, we shall do well to remove the short-time fluctuations from the field. But, if we are interested in short-time oscillations only, we should eliminate all long-time changes from our series.

Let us first be concerned with the study of long-time variations, or which comes to the same thing, trend. Since *the value of an item on a particular date in a time series consists of the long-time plus the short-time changes,* we can get a measure of the long-time change if we eliminate short-time changes from the series.

MEASUREMENT OF SECULAR TREND

If we plot our series on a graph paper, we shall observe an up and down movement in the curve. The index historigram of the retail prices of wheat drawn in Figure 28, is not a smooth curve: It is irregular. If we can smooth out these irregularities, the short-time oscillations shall be removed, and we shall obtain a measure of the trend values. Any of the following methods may be used for such smoothing, that is, for measuring the secular trend:

(1) Freehand Curve Method.

(2) Method of Moving Averages.

(3) Method of Least-Squares.

Freehand Curve Method.

We may observe the up and down movement of the curve and smooth out the irregularities by drawing a freehand curve or line through the index historigram such that the curve so drawn would give a general notion of the direction of the change. This freehand curve would eliminate the short-time oscillations and show the long-period general tendency of the retail price of wheat. This is exactly what is meant by trend.

But this method has a serious disadvantage that different persons may draw the freehand line at different positions and with different slopes. There is, therefore, the danger of different conclusions being drawn by different persons.

Method of Moving Averages.

It is an alternative method of ridding the historigram of its short period fluctuations, that is, of drawing a smooth curve. The procedure involved in calculating the moving averages has already been discussed in Chapter IX on Statis-

tical Averages (See pp. 139-140). To compute three-yearly moving average, for instance, the values of the 1st, 2nd and 3rd years are added up, the sum is divided by 3, and the quotient is placed against the 2nd year ; then, values of the 2nd, 3rd and 4th years are added up, averaged, and the average is placed against the 3rd year ; and, so on. The moving averages series would yield a smooth curve, provided the time interval chosen for computing the moving average is appropriate.

The question that, therefore, presents itself is as to what period of time should be used in calculating the moving averages. This question is of paramount importance since, no one period of time used in calculating the moving averages would eliminate short-time changes of *all* time series. For each series, the period of time in question will be decided in the light of its 'periodicity', for only the moving averages calculated by using that time-interval which approximately coincides with the 'periodicity' of a historigram, would eliminate, *nearly* completely, all its regular fluctuations and show the trend. This is the chief characteristic, as also the merit, of the Method of Moving Averages. In using this method, therefore, we must first search for 'periodicity' in the given series.

What is 'periodicity' and how is it measured? Periodicity is but another name for the average duration for which cycles in a series last. A cycle is said to be completed when beginning with a peak the falling curve reaches a minimum point and then rising again reaches the next peak, or, when beginning with a trough the rising curve reaches a maximum point and then falling again reaches the next trough. There may be several peaks and troughs in a given series. The simple average of the time-distances between the consecutive peaks and between the successive troughs of the waves of the historigram would give the approximate wave-length, the desired periodicity.

In Figure 28 the index-historigram is undoubtedly irregular in its shape, but the ups and downs show remarkable regularity in their occurrence. It will be observed that it shows recurring years of high prices at intervals of about 10 years, prominent crests falling in the years 1879, 1888, 1897 and 1908, and similarly recurring years of low prices again at intervals of about 10 years, prominent troughs falling in the

years 1876, 1885, 1894 and 1904. The average of these time-distances is 10 years, and it may now be said that the 'periodicity' of the given series is ten years, that is, the ups and downs repeat themselves in a cycle of nearly 10 years with marked regularity. If, therefore, we take 10 years as the time-

Retail price of wheat in India showing annual fluctuations, five-yearly moving average, and the trend.

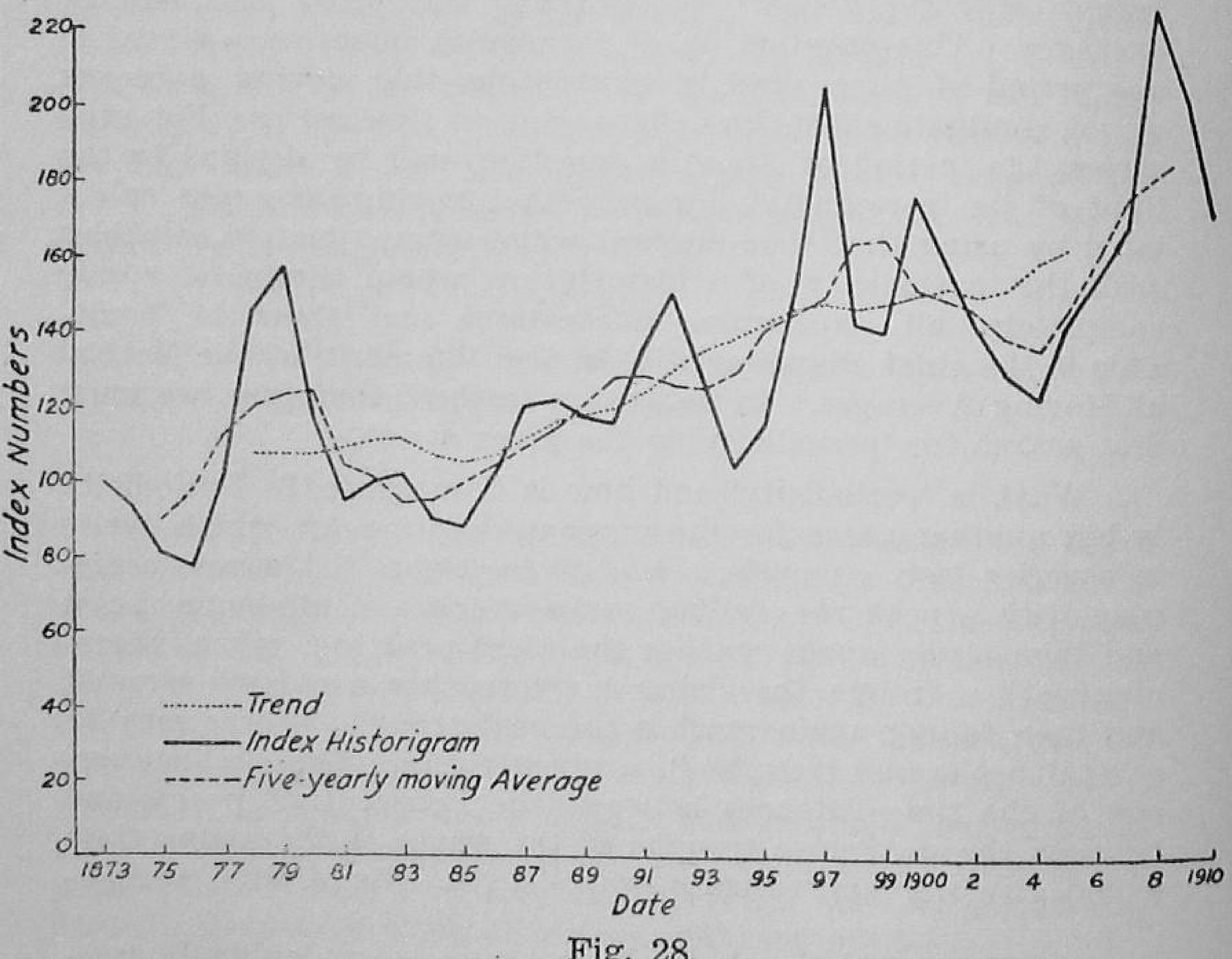


Fig. 28.

interval of the moving averages, the resulting series of moving averages will no longer show regular ups and downs; when plotted on graph paper, it will give a smooth curve; it is, in fact, the old series with short-period fluctuations eliminated; it gives the secular trend.

Ten is an even number and, therefore, the ten-yearly moving average can be placed only against the middle of the fifth and the sixth years of each group, as in column 4 of Table 51. We 'centre' these ten-yearly moving averages by taking two-yearly moving average of the figures given in column 4. For example, the ten-yearly moving average (centred), shown in column 5 of the Table, for the year 1881 is:

$$\frac{108+111}{2}=110, \text{ and for 1887 it is } \frac{110+115}{2}=113.$$

If, however, the periodicity were an odd number of years, there would have been no necessity to 'centre' the moving averages. The curve of the ten-yearly moving average (centred), drawn through the historigram in Figure 28, is far more smooth than the annual index-historigram or the five-yearly moving average curve. This curve shows the general rising tendency of wheat prices during the period under study. Column (5) in Table 51 gives the trend values. We have thus measured the trend by eliminating from the original series the short-time fluctuations, with the help of the method of moving averages.

Theory of Moving Averages Method.

It has been stressed above that the time-interval for computing the moving averages should coincide with the 'periodicity' of the given series. Periodicity, in the case of annual data, refers to cyclical fluctuations only; that is, to the period at which the cycle in the series repeats itself. Cyclical fluctuations, however, constitute only one of the three components of an item on any given data in a time series, inasmuch as an item in the time series is a summation of trend *plus* regular fluctuations *plus* irregular fluctuations. How is it then, that moving averages whose time-interval equals only the 'periodicity' of the series, should produce a smooth curve? How does this time-interval smooth out the irregular fluctuations and how does it safeguard that it will not distort the trend? The answers to these questions are to be found in the principles underlying the theory of moving averages as applied to time series analysis. These principles are described and illustrated below:—

(i) If the process of moving averages is used on a series which does not contain fluctuations at all, that is, on a series which is a trend series only; and, further, if the trend series is such as would be represented by means of a straight line when plotted, the moving averages will be exactly the same as the original series irrespective of the time-interval chosen for computing the moving averages.

Consider the series in Col. (2) of Table 51A. Whatever period is taken for striking the moving averages of this series, the original series will be exactly reproduced; that is, the moving average curve will coincide with the original series.

(ii) If the process of moving averages is used on a trend series and the trend series is such as would be represented by a curve when plotted, the process will reduce the curvature but the greater the time-interval chosen for the moving average, the more remote will be the moving average curve from the original.

Consider the series in Col. (3) of Table 51A. The five-yearly moving average curve of this series will be remote from the curve of original series, and the seven-yearly moving average curve will be remoter still.

(iii) If the moving average process is used on a series which consists of regular (cyclical or seasonal) fluctuations only, the time-interval which equals the periodicity of the series or is a multiple thereof, will eliminate the regular fluctuations entirely. Any other time-interval will only reduce, and not eliminate, them.

Consider the series in Col. (4) of Table 51A. The regular fluctuations repeat themselves every five years. Five- or ten-yearly moving averages will result in a zero series, three- or seven-yearly moving averages will produce series which will still contain fluctuations, though their magnitude may be less than that of the original series.

(iv) If the moving average process is used on a series which consists of irregular fluctuations only, no particular time-interval of the moving averages will eliminate them completely; but, the larger the time-interval, the greater will be the reduction in the fluctuations of the original series.

Consider Col. (5) of Table 51A. If three, five and ten-yearly moving averages are taken, each successive series will be smaller than its predecessor but none of them will be zero.*

Table 51A. *Trend, Regular and Irregular Fluctuations series.*

Year	Linear Trend Series	Curved Trend Series	Regular Fluctuations (Cyclical) Series	Irregular Fluctuations Series
(1)	(2)	(3)	(4)	(5)
1	2	2	+2	+1
2	4	4	+1	+2
3	6	7	0	−3
4	8	11	−1	−2
5	10	16	−2	+2
6	12	22	+2	0
7	14	29	+1	+1
8	16	37	0	+2
9	18	46	−1	0
10	20	56	−2	−1
11	22	67	+2	+2
12	24	79	+1	+1
13	26	92	0	0
14	28	106	−1	−1
15	30	121	−2	0
16	32	137	+2	−1

* To economic space, moving averages of the illustrative series given in Table 51A have not been worked out, nor have their graphs been drawn. For a clearer grasp of the four principles enunciated above, one would do well to compute the moving averages and draw the curves.

Now, an actual time series that we may have to analyse will in most cases be a combination of the types of series given in cols. (3), (4) and (5) in Table 51A. If, therefore, the time-interval used is too large, it will no doubt help to reduce the irregular fluctuations in the given series, but may distort the trend and may not completely eliminate the regular fluctuations. If, on the other hand, the time-interval used is too small, it may not distort the trend but may leave quite large irregular fluctuations and may not also remove the regular fluctuations. A middle course is, therefore, adopted in practice: that time-interval is used in the averaging process which will eliminate the regular fluctuations. It is hoped that the time-interval, so chosen, will also reduce the irregular fluctuations without introducing too much distortion in the trend. If this hope materialises well and good; if it does not, it only shows that the compromise between the time-intervals theoretically suitable for each of the components of the time series has not proved ideal in practice. It is, however, almost certain that no other compromise would produce a more smooth curve.

In Figure 28, the trend curve is not perfectly smooth. It does show some, although insignificant, ups and downs. Let us operate on our series, given in Table 51, with moving averages of other than ten years and observe if any of them smoothes out the fluctuations better. We begin five-yearly moving average, given in col. (3) of Table 51. This series of moving averages is plotted over the historigram in Figure 28. What do we see? The five-yearly moving average curve contains considerable fluctuations, though they are less than those in the historigram of annual indices. If we similarly plot seven, nine or eleven-yearly moving averages, we will still find somewhat greater fluctuations in the moving average curves than those shown by the ten-yearly one.

Limitations of the Method of Moving Averages.

It has been shown above that when we wish to use the method of moving averages, we search for periodicity in our series. The chief object of using this method being the elimination of periodic variations, the method is useful only in those time series which manifest more or less periodicity. The

method cannot be applied with equal success to any and every historigram. It is capable of being used where the series shows periodic fluctuations.

Another limitation of the method is that it does not enable the carrying out of the accurate trend to the extremes of the data. Trend values relating to the years from 1873 to 1877 and to those from 1906 to 1910 could not be ascertained in our example, in Table 51. This deficiency may, however, be made good in either of the following ways:—

(a) The moving average curve may be stretched freehand to each extreme.

(b) Artificial final groups may be formed by duplicating the numbers at the extremes. In Table 51, for instance, 170 might be added at the close of the given series five successive times, forming the required new groups for computing 10-yearly moving average up to 1910.

Both these ways are, of course, approximations.

The method of moving averages is, however, easy of application. It enjoys an advantage over the freehand curve method in that different people will not obtain different results by using it.

Smooth Curve and Short-time Changes.

The smooth curve is of no use, if one is interested in studying short-time changes. It shows the trend, the course that would be taken by a curve in the absence of disturbing factors, such as cyclical fluctuations. From the smoothed curve in Figure 28 we cannot study when prices began to rise or to fall. We cannot say that the retail price of wheat began to rise in 1887 because the smoothed curve begins to rise in that year. The average for 1887 is based on the prices of ten years, of which the year in question is only one. *To study short-time (annual) changes, that is, to study when prices began to rise and to fall, we must consult the original historigram, and not the trend.* It we study them from the trend, misleading conclusions might be drawn. For example, the shape of the smoothed curve for the period 1878 to 1880 (Figure 28) might lead one to think that the price of wheat was fairly steady

during that period, whereas according to the original data the indices 147, 158, and 118 for the years 1878, 1879 and 1880, respectively, show violent fluctuations. The smoothed curve, therefore, is good only for a study of long-period general tendency of the phenomenon under investigation.

Method of Least-Squares.

Although somewhat involved for practical use this is the most representative method. By using it, a *straight line trend* can be fitted to the given data. The trend line, technically called the line of "best fit," drawn under the least square assumption is the line from which the sum of the squares of the items, measured parallel to the y-axis, is the least. The following table illustrates the procedure to be adopted:

Table 51B. *Computation of Trend of the Sterling Assets of the Reserve Bank of India by the Method of Least-Squares.*

Year (1)	Sterling assets (crores of Rs.) (2)	Time deviation from middle year (3)	Deviation of col. (3) squared (4)	Product of col. (2) and col. (3) (5)	Trend ordinates (crores of Rs.) (6)
1936-37	83	−2	4	−166	67
1937-38	92	−1	1	− 92	84
1938-39	71	0	0	0	101
1939-40	90	+1	1	+ 90	118
1940-41	169	+2	4	+338	135
	505		10	+170	505

The procedure may be explained as follows:

(1) Compute the simple arithmetic average of the items. It is 505÷5=101 crores of Rupees. This average is the mid-point of the line of "best fit."

(2) From the mid-year of the series, 1938-39, find the time-deviations of the other years as in col. (3).

(3) Square the deviations, and add the resulting figures, as in col. (4).

(4) Multiply the value of the items by the time-deviations, and take the algebraic sum of the products, as in col. (5).

(5) Divide the sum of col. (5) by that of col. (4). The quotient, Rs. 17 crores, is the average increment in the trend year by year. It gives the slope of the line of "best fit."

(6) To obtain the trend ordinates shown in col. (6), put down the arithmetic average of the items, Rs. 101 crores, against the mid-year, 1938-39, and for the years after the mid-year add as many increments to the trend ordinate of the mid-year as the given year is away from the mid-year, while for the years before the mid-year subtract the increments in a similar manner. Thus, the trend ordinates for 1939-40 and 1940-41 are, respectively, (1×Rs. 17 crores)=Rs. 17 crores, and (2×Rs. 17 crores)=Rs. 34 crores more than the average, while the trend ordinates for 1937-38 and 1936-37 are, respectively, (1×Rs. 17 crores)=Rs. 17 crores, and (2×Rs. 17 crores)=Rs. 34 crores below the average. The algebraic sign of the increment must be kept in view while computing trend values.

Mathematical Treatment of the Method of Least Squares.

The procedure, explained above, of fitting a straight line trend to a given series can also be proceeded with mathematically. A straight line is a curve with curvature zero. The equation of a straight line is:

$$y=a+bx$$

In this equation, x is the independent variable, and y the dependent variable. In other words, the variation of y is determined by the variation of x ; or, y is the 'function' of x. Also, a and b are two constants in the equation, of which b indicates the slope of the line. Therefore, to fit a straight line trend to any given series, we have to frame the above equation.

To estimate the value of the constants in the above equation, we start with normal equations. The following are the two normal equations for fitting a straight line by the method of least squares:—

$$\Sigma(y)=na+b\Sigma(x) \quad \ldots\ldots\ldots\ldots (1)$$

$$\Sigma(xy)=a\Sigma(x)+b\Sigma(x)^2 \quad \ldots\ldots\ldots\ldots (2)$$

We apply these to the example already worked out in Table 51B, above. The x variable in this case is time, and the y variable the amount of sterling assets. We may number the x's serially from 1 to 5, and proceed as in Table 51C.

Table 51C. *Computation of Straight Line Trend of the Sterling Assets of the Reserve Bank of India by the Method of Least Squares.*

Year	Sterling assets (crores of Rs.) observed y	x	xy	x^2	Computed y (crores of Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1936-37	83	1	83	1	67
1937-38	92	2	184	4	84
1938-39	71	3	213	9	101
1939-40	90	4	360	16	118
1940-41	169	5	845	25	135
Total	505	15	1685	55	505

From the above table, it will be seen that,

$$n = 5$$
$$\Sigma(y) = 505$$
$$\Sigma(xy) = 1685$$
$$\Sigma x = 15$$
$$\Sigma x^2 = 55$$

The normal equations are :

$$\Sigma(y) = n(a) + b\,\Sigma(x) \quad \ldots\ldots\ldots\ldots\ldots (1)$$

$$\Sigma(xy) = a\,\Sigma(x) + b\,\Sigma(x)^2 \quad \ldots\ldots\ldots\ldots\ldots (2)$$

Substituting the above values in the normal equations, we have,

$$505 = 5a + 15b \quad \ldots\ldots\ldots\ldots\ldots (1)$$

$$1685 = 15a + 55b \quad \ldots\ldots\ldots\ldots\ldots (2)$$

We solve the above as simultaneous equations for finding out the values of the constants a and b. Multiplying (1) by 3, we have,

$$1515 = 15a + 45b \quad \ldots\ldots\ldots\ldots\ldots (1)$$
$$1685 = 15a + 55b \quad \ldots\ldots\ldots\ldots\ldots (2)$$
$$170 = 10b \quad \text{by subtracting (1) from (2)}$$
$$\text{or, } b = 17$$

To find the value of a, we substitute the value of b in equation (2). We have,

$$1685 = 15a + 935$$
$$\text{or, } 15a = 750$$
$$\text{or, } a = 50$$

The equation of straight line, as already said, is:

$$y = a + bx$$

Substituting in the above equation, the calculated values of a and b, we have,

$$y = 50 + 17x$$

This is the equation of the straight line trend for the data given in Table 51C.

To find y for different values of x (from 1 to 5), we multiply the value of b by 1, 2, 3, etc., and add to the product the value of a. Thus,

$$\text{when } x=1,\ y=50+17=67$$
$$x=2,\ y=50+34=84$$
$$x=3,\ y=50+51=101$$
$$x=4,\ y=50+68=118$$
$$x=5,\ y=50+85=135$$

The trend ordinates, as calculated above and shown in col. (6) of Table 51C, are the same as those in col. (6) of Table 51B. The two approaches have resulted in the same values of trend ordinates. Figure 28A shows the original data and the straight line trend.

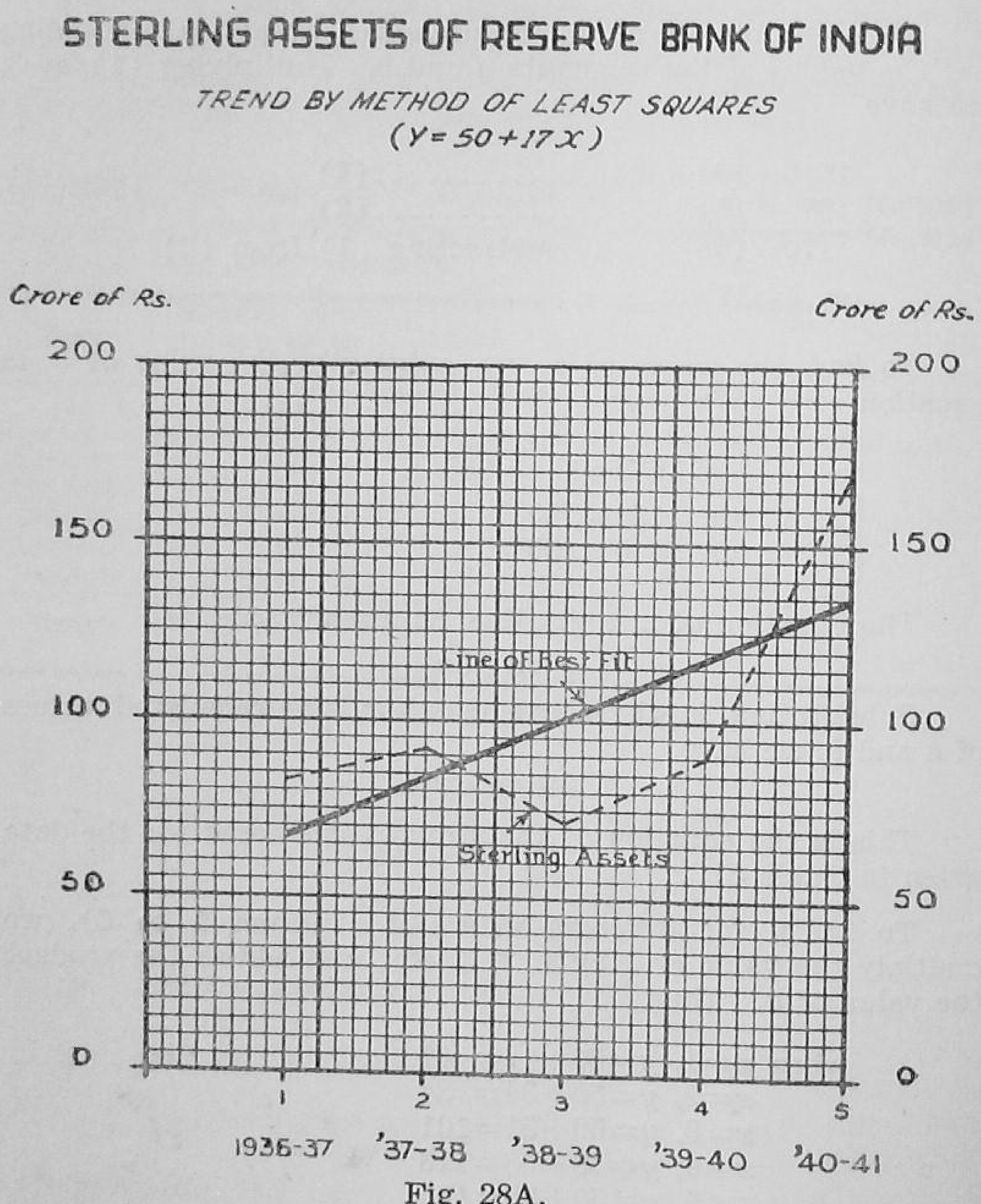

STERLING ASSETS OF RESERVE BANK OF INDIA
TREND BY METHOD OF LEAST SQUARES
(Y=50+17X)
Crore of Rs.
Crore of Rs.
200
150
100
50
0
Line of Best Fit
Sterling Assets
1
2
3
4
5
1936-37
'37-38
'38-39
'39-40
'40-41

Fig. 28A.

Characteristics and Limitation of Line of Best Fit.

The values of col. (6) in Table 51B or 51C have been plotted in Fig. 28A along with original data. The thick curve shows the straight line trend of the sterling assets of the Reserve Bank of India. It should be clearly noted that this trend line, the line of "best fit," passes through the arithmetic means of the x and the y series, x series representing time and y the sterling assets. The average of x in our example is 3; that is, the year 1938-39. The average of y is 101. It will be seen from Fig. 28A that the line of "best fit" passes through both these ordinates, 3 and 101. This line, it should be further noted, is the "best fit" only for the period to which it has reference: the addition of sterling assets for a few more years, or the elimination of some of those already included, may change the position materially.

Parabolic Trend.

In economic statistics, a straight line trend is not so common as the parabolic. Economic phenomena can be represented by a parabola of the second or the third order. The equation for the curve of a parabola of the second order is:

$$y=a+bx+cx^2$$

The normal equations used for fitting a parabola of the second order are:

$$\Sigma(y)=na+b\,\Sigma(x)+c\,\Sigma(x^2)\ldots\ldots\ldots\ldots(1)$$
$$\Sigma(xy)=a\,\Sigma(x)+b\,\Sigma(x^2)+c\,\Sigma(x^3)\ldots\ldots\ldots\ldots(2)$$
$$\Sigma(x^2y)=a\,\Sigma(x^2)+b\,\Sigma(x^3)+c\,\Sigma(x^4)\ldots\ldots\ldots\ldots(3)$$

The procedure for computing the values of the constants is the same as already shown in Table 51C; only additional columns will have to be provided for calculating x^2y, x^3 and x^4. That is simple to do, once one has carefully studied Table 51C.

MEASUREMENT OF SHORT-TIME FLUCTUATIONS

We are very often interested in the study of short-time oscillations of a time series. For this study we should eliminate the trend, or long-time variations, from the given series.

To do so is quite easy when once the trend of the series has been known; for, the difference between the value of an item on a particular date and its corresponding trend value is the short-time oscillation. These differences, called deviations, are taken with their proper algebraic signs, and, when plotted on a horizontal base line show the short-time oscillations.

In Table 51, we have calculated the trend values in column

Retail price of wheat in India showing short-time oscillations (with the trend removed.)

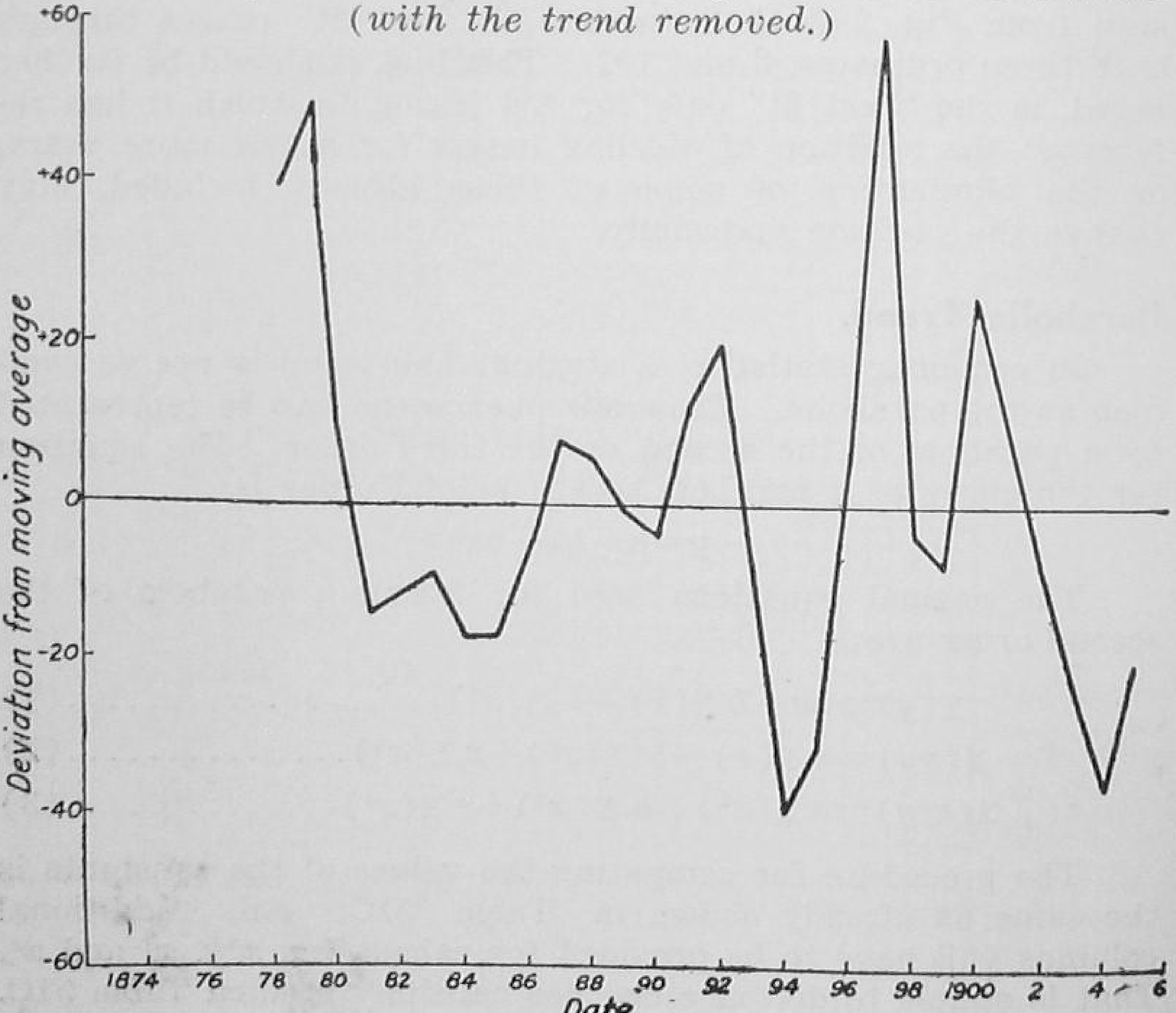


Fig. 29.

5. To eliminate the trend, we compute deviations of the prices from the trend values. These deviations are placed in column 6 of the same table. Column (6) now gives the short-time fluctuations of retail prices of wheat in India. They are graphi-

cally presented in Figure 29. The curve in the Figure reveals the short-time oscillations, which in the index historigram in Figure 28 had been obscured by the long-time variations. The short-time oscillations would, it should be noted, consist of regular (seasonal and cyclical) and irregular fluctuations which may also be separately measured.

MEASUREMENT OF SEASONAL VARIATIONS

So far, in our examples, we considered only annual data. Annual data are good enough for studying the effect of cyclical factor. Seasonal factor recurs at much shorter intervals than cycles. Therefore, to study the effect of seasonal factor in a series, we should have weekly or monthly quantities. If monthly records of some data are available over a series of years, either of the following methods may be used to arrive at a measure of their normal seasonal variation:

(1) Computation of Seasonal Average.

(2) Computation of Seasonal Index by using monthly averages.

(3) Computation of Seasonal Variations by using moving averages.

Seasonal Average.

Seasonal average is obtained by adding up the amounts separately for each month of the year, that is, totalling the amounts for the Januaries, the Februaries, etc., separately, and dividing each such summation by the total number of years for which the data are given. In Table 52, column 8 gives the seasonal average for the exports of raw jute from undivided India between 1939 and 1943.

Seasonal Variation Index.

To obtain the seasonal variation index, find out the average of monthly averages, and then find the percentage that each monthly average bears to the average of monthly averages. In Table 52, the 'average of monthly averages', given at the

bottom of column (8), is 28,000 tons. Therefore, the seasonal index for January is

$$\frac{42200 \times 100}{28000} = 150.7$$

Table 52. *Calculation of Seasonal Average and Seasonal Variation Index of the Exports of Raw Jute from undivided India.*

Months	Raw Jute					Total of five years	Five years' average	Percentage of Monthly total to average Monthly total
	1939	1940	1941	1942	1943			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Tons (000)	Tons (000)	Tons (000)	Tons (000)	Tons (000)	Tons (000)	Tons (000)	
Jan.	62	59	35	33	22	211	42.2	150.7
Feb.	69	68	17	15	14	183	36.6	130.7
March	60	51	10	24	13	158	31.6	112.9
April	53	38	20	26	17	154	30.8	110.0
May	44	36	31	6	12	129	25.8	92.2
June	34	16	36	16	12	114	22.8	81.4
July	21	8	26	28	13	96	19.2	68.6
Aug.	37	11	30	14	8	100	20.0	71.4
Sept.	45	11	23	22	11	112	22.4	80.0
Oct.	31	17	29	21	15	113	22.6	80.7
Nov.	65	12	30	38	22	167	33.4	119.3
Dec.	60	22	26	22	13	143	28.6	102.1
Total						1680	336.0	1200.0
Average						140	28	100

Seasonal indices can also be calculated directly from the 'total of five years', in column (7). Thus, the seasonal index for January is

$$\frac{211000 \times 100}{140000} = 150.7$$

Both the approaches will give the same indices (col. 9).

The seasonal average in column (8) and the seasonal index in column (9) show the typical movement of exports of raw jute. If both of them are plotted on graphs, the part of the year in which the exports are the greatest and the part in which they are the least would become easily visible. The influence of the seasonal factor on the exports would thus be rendered clear. The normal seasonal variation, thus computed, would recur month by month in the same manner year after year.

In Table 52 we have pooled together the experience of only five years. In practice a longer period should be used.

Seasonal Variation.

If monthly quantities are available for a number of years, 12-monthly moving averages of the series will give the trend values. Since 12 is an even number, the moving averages will have to be 'centred', as was done in Table 51, col. 5. Original data *minus* the trend will give the short-time fluctuations. These fluctuations may then be averaged, like the seasonal average in col. 8 of Table 52. The resulting series of averages will provide a measure of seasonal variations.

Table 52A gives the monthly output of coal in India during July 1948 to June 1951 in col. (2), the 12-monthly moving average (centred) in col. (3)* and the short-time fluctuations in col. (4). These short-time oscillations are averaged, as in Table 52B.

---

*Moving averages (centred) in col. (3) of Table 52A are based on the data from January, 1948 to December 1951. They have first been calculated, as in col. (4) of Table 51, and then centred by averaging in two's, as in col. (5) of Table 51. In order to compress Table 52A into one page, production data for January-June, 1948 and July-December, 1951, have been omitted from col. (2).

Table 52A. *Time Analysis of Coal Production in India.*

(Lakh tons)

Year/Month	Production	12-monthly moving average (centred)	Short-time fluctuations	Seasonal variations	Irregular fluctuations
(1)	(2)	(3)	(4)	(5)	(6)
1948					
July	22.3	24.9	(—)2.6	(—)3.0	(+)0.4
Aug.	22.3	24.9	(—)2.6	(—)1.2	(—)1.4
Sept.	25.6	24.9	(+)0.7	(+)0.9	(—)0.2
Oct.	24.6	24.9	(—)0.3	(—)0.6	(+)0.3
Nov.	22.2	25.0	(—)2.8	(—)1.7	(—)1.1
Dec.	25.7	25.0	(+)0.7	(+)1.2	(—)0.5
1949					
Jan.	25.7	25.0	(+)0.7	(+)0.5	(+)0.2
Feb.	27.1	25.3	(+)1.8	(+)1.9	(—)0.1
March	27.0	25.5	(+)1.5	(+)1.8	(—)0.3
April	27.0	25.6	(+)1.4	(+)0.8	(+)0.6
May	26.0	25.8	(+)0.2	(+)0.6	(—)0.4
June	23.7	26.1	(—)2.4	(—)2.2	(—)0.2
July	23.8	26.3	(—)2.5	(—)3.0	(+)0.5
Aug.	26.3	26.4	(—)0.1	(—)1.2	(+)1.1
Sept.	27.3	26.5	(+)0.8	(+)0.9	(—)0.1
Oct.	25.9	26.6	(—)0.7	(—)0.6	(—)0.1
Nov.	26.0	26.7	(—)0.7	(—)1.7	(+)1.0
Dec.	28.8	26.8	(+)2.0	(+)1.2	(+)0.8
1950					
Jan.	26.1	26.8	(—)0.7	(+)0.5	(—)1.2
Feb.	29.4	26.7	(+)2.7	(+)1.9	(+)0.8
March	28.8	26.7	(+)2.1	(+)1.8	(+)0.3
April	27.2	26.8	(+)0.4	(+)0.8	(—)0.4
May	27.1	26.8	(+)0.3	(+)0.6	(—)0.3
June	24.4	26.7	(—)2.3	(—)2.2	(—)0.1
July	23.0	26.8	(—)3.8	(—)3.0	(—)0.8
Aug.	26.0	26.9	(—)0.9	(—)1.2	(+)0.3
Sept.	28.2	27.0	(+)1.2	(+)0.9	(+)0.3
Oct.	26.2	27.1	(—)0.9	(—)0.6	(—)0.3
Nov.	25.6	27.2	(—)1.6	(—)1.7	(+)0.1
Dec.	28.2	27.4	(+)0.8	(+)1.2	(—)0.4
1951					
Jan.	29.1	27.6	(+)1.5	(+)0.5	(+)1.0
Feb.	28.9	27.8	(+)1.1	(+)1.9	(—)0.8
March	29.6	27.9	(+)1.7	(+)1.8	(—)0.1
April	28.5	28.0	(+)0.5	(+)0.8	(—)0.3
May	29.7	28.3	(+)1.4	(+)0.6	(+)0.8
June	26.7	28.5	(—)1.8	(—)2.2	(+)0.4

Table 52B. *Computation of Seasonal Variations.*

(Lakh tons)

Month	Deviations from Trend				Seasonal variations
	1948	1949	1950	1951	(average)
January	—	(+)0.7	(−)0.7	(+)1.5	(+)0.5
February	—	(+)1.8	(+)2.7	(+)1.1	(+)1.9
March	—	(+)1.5	(+)2.1	(+)1.7	(+)1.8
April	—	(+)1.4	(+)0.4	(+)0.5	(+)0.8
May	—	(+)0.2	(+)0.3	(+)1.4	(+)0.6
June	—	(−)2.4	(−)2.3	(−)1.8	(−)2.2
July	(−)2.6	(−)2.5	(−)3.8	—	(−)3.0
August	(−)2.6	(−)0.1	(−)0.9	—	(−)1.2
September	(+)0.7	(+)0.8	(+)1.2	—	(+)0.9
October	(−)0.3	(−)0.7	(−)0.9	—	(−)0.6
November	(−)2.8	(−)0.7	(−)1.6	—	(−)1.7
December	(+)0.7	(+)2.0	(+)0.8	—	(+)1.2

The last column in Table 52B gives the seasonal variations. They have been copied out from that table into col. (5) of Table 52A. It will be seen that generally the seasonal factor tends to push up coal production to its peak in February each year and to drag it down to the bottom in July after the rainy season sets in. Study Fig. 29A.

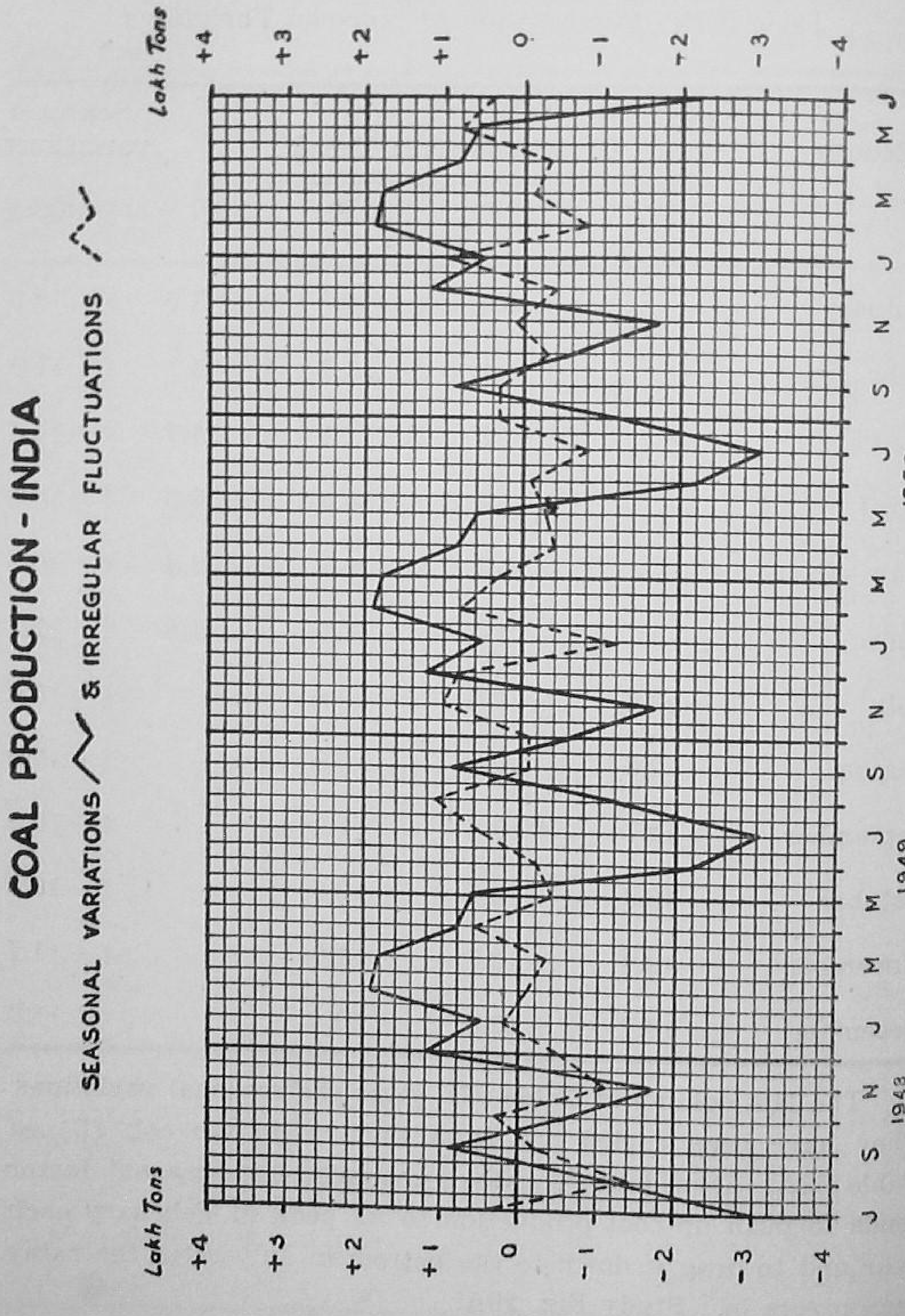

COAL PRODUCTION - INDIA
SEASONAL VARIATIONS & IRREGULAR FLUCTUATIONS
Lakh Tons
+4
+3
+2
+1
0
-1
-2
-3
-4
J S N J M M J S N J M M J S N J M M J
1948
1949
1950
1951

Fig. 29A.

MEASUREMENT OF IRREGULAR FLUCTUATIONS

If short-time fluctuations comprise seasonal and irregular variations, it is simple to measure the irregular fluctuations once the seasonal variations have been measured. In Table 52A, col. (4) gives the short-time fluctuations and col. (5) the seasonal variations. Irregular fluctuations too are shown in that table, in col. (6). They have been arrived at by subtracting col. (5) from col. (4), that is, by eliminating seasonal variations from short-time fluctuations.

Seasonal variations and irregular fluctuations have been plotted in Fig. 29A. It will be seen from the Figure that the irregular fluctuations are adventitious and that unlike seasonal variations positive and negative signs in them occur at random. There is also no fixed or definable relationship between the nature of variation of the seasonal and irregular fluctuations: both may be positive or negative, or run in opposite directions. Even where they vary in the same direction, the magnitude of irregular fluctuations bears no definite relationship to seasonal variations: though in most cases the irregular fluctuations are of smaller size than the regular ones, yet the opposite is also true in several cases. Figure 29A, therefore, clearly emphasises that while there is a definite regularity about the seasonal variations, there is lack of it in so far as the irregular fluctuations are concerned.

COMPARISON OF TIME CHANGES

We have so far answered only the first of the two questions with which this Chapter opened. We may now turn to the second: how are changes over a period of time in one series related to those in another series?

When comparison of time-changes in two historical variables is desired, it would be well to first convert them into index numbers, so that their relative fluctuations may be easily compared. The trends of the two index series may then be measured and plotted on one graph with the same base and scale. The direction of the movement of the two curves, and the magnitude of movement may then be studied from the graph. Likewise, the short-time oscillations of the two index series may be compared by plotting them on another graph.

If a quantitative measure of the degree of relationship between the long-period or short-period fluctuations of two time series is desired, methods of correlation, which will be discussed in the next Chapter, will have to be used.

UTILITY OF TIME SERIES ANALYSIS

Table 52A demonstrates how a time series can be broken up into long-period and short-period fluctuations with the help of the method of moving averages, and how the short-time fluctuations can be further split into their constituents—the regular and the irregular variations. This time series analysis is of great importance in economic statistics. It is useful to know not only the fluctuations in production, prices, sales, costs, etc., of the current time, but also what their general movement has been over a number of years. A knowledge of the period at which business cycles have occurred in the past, as also that of seasonal oscillations in a trade or industry, is equally useful. All this information would enable a businessman to effect adjustments in his business as and when necessary rather than be caught unawares, an economist to foresee what is likely to come and an administrator to mould his policies. A close and regular watch on the results of statistical analysis is sure to mitigate the difficulties of practical businessmen when the turning point arrives. Past experience is a guide to the future; time series analysis only brings to light the salient features of this past experience.

EXERCISES

(1) (*a*) Distinguish between the regular and irregular fluctuations in a Time Series.

(*b*) Write a short note on the value of analysing time variations.

(M.A., Punjab, Apr., 1952).

(2) Write a short essay on "Analysis of Time Series."

(B. Com., Hons., Andhra, 1943).

(M.A., Patna, 1944).

(3) Explain clearly what is meant by time series analysis. Indicate fully the importance of such analysis in business.

(B. Com., Luck., 1944).

(4) Describe briefly the statistical procedure you would adopt for the analysis of time series, and explain carefully how you would isolate the secular trend.

(M.A., Patna, 1942).

(5) Indicate briefly how you would analyse a series of monthly records extending over 50 years.

(M.A., Alld., 1942).

(6) What is meant by 'trend'? How would you statistically eliminate the influence of seasonal and cyclic factors on the long-period movement of any series?

(B. Com., Bombay, 1936).

(7) Discuss the claims and limitations of the method of moving averages as applied to analysis of time series.

(M.A., Delhi, 1953).

(8) Explain and illustrate the principles underlying the method of moving averages as applied to analysis of time series, and show why one has to search for 'periodicity' in a given series while proceeding to analyse it.

(9) (*a*) Explain fully what is meant by secular trend, seasonal variations, and cyclical fluctuations, illustrating your answer.

(*b*) Study the short-time fluctuations of the following temperatures measured in degrees Fahrenheit:—

Date 1941		Temperature	Date 1941		Temperature
Feb.	1	40	Feb.	11	78
,,	2	50	,,	12	80
,,	3	44	,,	13	60
,,	4	70	,,	14	64
,,	5	52	,,	15	62
,,	6	44	,,	16	68
,,	7	36	,,	17	86
,,	8	40	,,	18	96
,,	9	56	,,	19	94
,,	10	68	,,	20	78

(B. Com., Alld., 1942).

(10) Explain fully what you understand by secular trend.

Assuming a ten-yearly cycle for the following series relat-

ing to Index Numbers of the Retail Price of Wheat in India (1873=100), give the trend values, and represent graphically the short-time fluctuations with the trend removed.

Year	Annual Average	Year	Annual Average	Year	Annual Average
1906	155	1914	200	1922	315
1907	168	1915	227	1923	356
1908	226	1916	193	1924	246
1909	203	1917	205	1925	294
1910	170	1918	270	1926	281
1911	153	1919	341	1927	267
1912	170	1920	310	1928	264
1913	177	1921	360	1929	262

(M. Com., Alld. 1944).

(11) Represent the following data graphically, and using 3-year moving averages, indicate the trend of the series on the graph:—

Year	Birth-Rate	Year	Birth-Rate	Year	Birth-Rate
1917	30.9	1924	31.0	1931	23.1
1918	30.2	1925	29.0	1932	23.7
1919	29.1	1926	27.9	1933	22.6
1920	31.4	1927	27.7	1934	23.6
1921	33.4	1928	26.4	1935	23.0
1922	30.2	1929	24.7	1936	22.0
1923	30.4	1930	24.1	1937	22.6
				1938	22.9

(M.A., Alld., 1951).

(12) The following table shows the available supplies of all cereals per adult in India for a number of years:

Year	lbs.	Year	lbs.
1931—32	427	1938—39	360
1932—33	416	1939—40	382
1933—34	408	1940—41	352
1934—35	408	1941—42	355
1935—36	379	1942—43	358
1936—37	405	1943—44	383
1937—38	390	1944—45	361

Draw a graph to represent the above data. Calculate 3-yearly moving averages and plot them on the same chart.

What are the deviations of the actual values from the moving average for the years 1941-42 and 1943-44?

(M. Com., Raj., 1951).

(13) Explain how you will deal with a time series and illustrate your remarks with the help of the following series:

Quarter	1923	1924	1925	1926	1927	1928	1929	1930
1	102	102	106	108	110	109	121	106
2	107	90	102	107	110	109	125	204
3	101	88	101	108	104	110	122	91
4	98	98	106	108	101	114	108	84

(B. Com., Nag. 1942).

(14) Analyse the following figures of output of Coal in Great Britain so as to arrive at the extent of (a) Seasonal Movement and (b) Irregular Fluctuations:—

Year	Quarters	Output (million tons)
1927	1	68.3
”	2	62.6
”	3	61.1
”	4	63.3
1928	1	65.4
”	2	57.9
”	3	56.4
”	4	61.5
1929	1	68.1
”	2	62.7
”	3	62.8
”	4	67.0
1930	1	70.1
”	2	59.1
”	3	56.3
”	4	61.6
1931	1	59.5
”	2	54.8
”	3	51.1
”	4	58.0

(M.A., Punjab, Sept., 1951).

(M. Com., Alld., 1947).

(15) (*a*) How would you distinguish the cyclical fluctuations from the trend and the seasonal fluctuations?

(*b*) The following table gives the value of the exports of merchandise from undivided India during the years 1919-20 to 1923-24. Calculate the seasonal variations for each month during this period.

Months		In Crores of Rupees				
		1919-20	1920-21	1921-22	1922-23	1923-24
April	..	20	27	17	23	29
May	..	20	26	18	26	28
June	..	19	21	15	18	29
July	..	26	19	17	23	25
August	..	25	19	18	24	22
September	..	30	21	19	20	23
October	..	28	18	17	21	25
November	..	29	17	19	27	26
December	..	26	18	21	26	30
January	..	29	18	22	28	36
February	..	26	17	21	30	35
March	..	30	18	26	31	40

(M.A., Alld., 1937).

(16) Calculate the seasonal variation index from the data given in exercise 15 (*b*) above.

(17) Measure the trend, seasonal variations and irregular fluctuations of the data given in exercise 15 (*b*) above. Plot the seasonal and irregular variations on the same graph.

(18) Classify the different types of fluctuations which occur in the analysis of time-series. Illustrate your remark with the help of the following series:

6099	6497	6898	7300	7699
6223	6621	7024	7421	7828
6351	6754	7152	7553	7949
6477	5878	7275	7675	8077

(M.A., Cal., 1937).
(B. Com., Nag., 1944).

(19) Write a note on the statistical analysis of time-series in economic studies. Illustrate your remarks with the help of the following table, using in particular 3-year and 5-year moving averages:—

Year	Value	Year	Value	Year	Value	Year	Value
1901	507	1908	552	1915	583	1923	628
1902	522	1909	556	1916	581	1924	632
1903	524	1910	548	1917	599	1925	626
1904	521	1911	572	1918	612	1926	644
1905	538	1912	569	1919	597	1927	643
1906	541	1913	567	1920	612	1928	642
1907	537	1914	587	1921	616	1929	661
				1922	608	1930	659

(M.A., Cal., 1935).

(20) The following are the Monthly Index Numbers of Commodity Group (Food and Tobacco) issued by the Economic Adviser to the Government of India:

Prices for week ending 19*th August* 1939=100

	Month	Index		Month	Index
1941			1942	July	155.8
	Oct.	127.4		Aug.	158.9
	Nov.	127.9		Sept.	161.0
	Dec.	127.5		Oct.	167.2
1942				Nov.	172.4
	Jan.	128.4		Dec.	178.5
	Feb.	132.3	1943		
	Mar.	130.5		Jan.	190.8
	Apr.	136.5		Feb.	270.0
	May	144.7			
	June	152.3			

Fit a straight line trend to the above data by the Method of the Least Squares and exhibit the graph of the data with the trend.

(M. Com., Luck., 1944).

(21) Plot the following figures on a graph paper and study their trend. On a separate graph paper show their short-time oscillations with the trend removed.

Year		Value	Year		Value
1913-14	..	264	1924-25	..	305
-15	..	255	-26	..	303
-16	..	267	-27	..	306
-17	..	267	-28	..	297
-18	..	269	-29	..	292
-19	..	264	-30	..	304
-20	..	263	-31	..	310
-21	..	255	-32	..	317
-22	..	271	-33	..	331
-23	..	289	-34	..	344
-24	..	310	..	..	..

(22) Fit straight line trends by the method of least squares to the following data spread over 10 consecutive years Also plot the data and the trend.

(1)	(2)	(3)
1	20	110
2	19	125
4	17	135
5	16	150
6	15	165
7	14	155
8	13	175
9	12	180
10	11	200

(23) Fit a straight line TREND by the method of LEAST SQUARES to the growth of the reserves of Co-operative Societies in India as given below, and plot the series as well as the TREND on graph paper.

Year	Reserves (Lakhs of Rs.)
1927—28	612
1928—29	719
1929—30	820
1930—31	907
1931—32	1,001
1932—33	1,106
1933—34	1,231

(M.A., Punjab, April, 1951).

(24) Convert the following yield of major foodgrains in India into index numbers with figures for 1938-39 equated to 100, and fit a straight-line trend by the method of least squares to the indices thus computed:—

Year	Yield (in millions of tons)
1934—35	52.7
1935—36	49.9
1936—37	55.5
1937—38	54.3
1938—39	49.6
1939—40	53.0
1940—41	50.2

(Index numbers and trend need not be graphically presented.)

(M. Com., Alld., 1945).

(25) Fit straight line trend, by the Method of Least Squares, to the following statistics of yield of sesamum in India:—

Year	Yield ('000)
1910—11	499
1911—12	367
1912—13	442
1913—14	378
1914—15	524
1915—16	460
1916—17	482
1917—18	363
1918—19	235
1919—20	225

(M. Com., Alld., 1948).

(26) In an experiment designed to find the effect of seed rate on the yield of wheat, the following results were obtained:—

Seed Rate (lbs. per acre) ..	40	50	60	70	80
Average Yield of Wheat (lbs. per acre) ..	850	862	888	817	768

Draw the graph. Fit the second degree parabola and sketch it.

(M.A., Raj., 1950).

(27) Draw a graph of the following time series and study its trend:—

Year	Value	Year	Value	Year	Value
1920	505	1927	734	1934	999
1921	610	1928	652	1935	883
1922	630	1929	750	1936	805
1923	670	1930	980	1937	900
1924	575	1931	820	1938	1,050
1925	680	1932	743	1939	935
1926	895	1933	798	1940	930

(B. Com., Alld., 1945).

(28) The number (in hundreds) of letters posted in a certain city on each day in a typical period of five weeks was as follows:

	Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.	Total for each week.
1st week	18	161	170	164	153	181	76	923
2nd week	18	165	179	157	168	195	85	968
3rd week	21	162	169	153	139	185	82	911
4th week	24	171	182	170	162	179	95	983
5th week	27	162	186	170	170	182	120	1,017
Total for all weeks	108	821	886	814	792	922	458	4,801

Calculate indices of the average fluctuations within the week for this period, and show how to eliminate the fluctuation from the data.

(B. Com., Andhra, 1942).

CHAPTER XV

CORRELATION

Black cats cause bad luck while filled-up pitchers good fortune—these are the beliefs held by some people. But these beliefs are incapable of being justified by any mathematical theory. It is, therefore, difficult to say if there really exists any relationship between black cats and bad luck and between filled-up pitchers and good fortune, though occasional coincidences may suggest such notions. On the other hand, some people believe that devaluation of the rupee from 1s. 6d. rate to 1s. 4d. would stimulate India's export trade, or that a rise in the rate of interest would encourage savings. These impressions do indicate some sort of relationship, but they are mere guesses until the relationship is established mathematically. The theory by means of which quantitative connections between two sets of phenomena are determined is called the Theory of Correlation.

Meaning of Correlation.

Correlation or co-variation is the connection, relationship or interdependence between two quantities that vary in sympathy. If in each of them some factor is numerically measured and it is discovered that large values of one go with large values of the other and small values with small, or *vice versa*, the two factors exhibit some mutual dependence, termed correlation. In other words, if two quantities vary in sympathy so that a movement—an increase or decrease—in the one tends to be accompanied by a movement in the same or inverse direction in the other, and the greater the volume of change in the one the greater is the volume of change in the other, the quantities are said to be correlated.

In natural sciences correlation can be reduced to absolute mathematical terms. Heat always increases with light and an electric current is always associated with magnetic field. These instances suggest a high degree of correlation. But in social sciences it is seldom that any absolutely fixed mathematical relationship between two variables can be established. The law of demand, the law of diminishing returns, Gresham's law,

to take a few illustrations, suggest correlation, but this correlation is not so perfect as that in the natural sciences. Therefore, *in inexact sciences we must take the fact of correlation established, if in a large number of cases two variables always tend to move in the same or opposite directions.*

Such phenomena are not uncommon in the social and economic sphere. We very often see that demand for a commodity generally falls with a rise in its price, that price level in a country generally rises with supply of money, that tall fathers generally have tall sons, that young husbands generally have young wives, that a taller man generally tends to be thinner. In all these cases correlation exists, but the relationship exists in general, or on the whole. It is not as if demand for every commodity would fall with a rise in its price; in some cases demand may even rise.

Correlation does not also imply the existence necessarily of direct causal relationship between the concerned variables. If price level in a country generally rises with supply of money, it does not necessarily prove that rise in price level is the effect of increase in money supply.

Positive and Negative Correlation.

Correlation may be positive or negative. If the two given variables steadily deviate in the same direction, correlation is direct or positive; but if they constantly deviate in the opposite directions, correlation is inverse or negative. That is, if an increase (or decrease) in the values of one variable is associated with an increase (or decrease) in the values of the other, correlation between them is positive. And, if an increase (or decrease) in the values of one variable is associated with a decrease (or increase) in the values of the other, correlation between them is negative.

Linear and Curvilinear Correlation.

When the amount of movement in one variable bears a constant ratio to the amount of change in the other, the correlation is said to be linear. In case the ratio of change is not constant, the correlation is called curvilinear or non-linear. Thus, the distinction between linear and curvilinear correlation depends upon the constancy of ratio change between two variables.

If the output of workers in a factory is doubled by doubling the number of workers, and varies in quantity as the number of workers is reduced or increased, there is linear correlation.

Study and Measurement of Correlation.

Correlation may be studied between (1) two *related* historical variables and (2) between any other two groups of *related* phenomena. Correlation may, for instance, be studied between production of cereals in India and their imports over a period of time to find whether the imports fall with an increase in production in the country; it may be studied between supply of a commodity and its price over a period of time to find whether the price falls with increase in supply. These are examples of related historical variables. Correlation may also be studied between the length and breadth of the leaves of a certain tree to find if longer leaves tend to be less broad, or between the statures of fathers and sons to find if tall fathers generally have tall sons. These are examples of non-historical but related phenomena. If, however, one produces figures to show that as the production of cane sugar increased in India that of motor cars fell in the U. S. A. over a period of time, it would not imply correlation, unless there is reason to believe that production of cane sugar in India and that of motor cars in the U. S. A. are really related in some way.

The following methods may be used for the study and measurement of correlation:

(1) Computing the co-efficient of correlation;
(2) Constructing a graph or a scatter diagram;
(3) Showing the series in a correlation table.

COEFFICIENT OF CORRELATION

Degrees of Correlation.

Correlation exists in various degrees. The radius of a circle bears a perfectly definite relationship to its area, so that the area increases in a perfectly definite proportion with an increase in the radius. Similarly, the area of a square increases in a definite ratio with an increase in the length of its side. These are the instances where correlation or co-variation is perfect and positive. Correlation will be perfectly negative, if

a fall of 10 per cent in the price of a commodity results in 10 per cent rise in its demand.

There may be instances where correlation may not exist: If the height of a house were compared with that of a growing tree over a period of time, it would be found that while the height of the house remained unchanged during the period, that of the tree kept on increasing. Evidently, the height of the house cannot be associated with that of the tree and, therefore, no correlation exists between them.

Correlation may exist in a limited degree. If demand for a commodity increases, its price also increases, but not necessarily in the same proportion. This is a case of limited positive correlation. If area under food crops in a country increases, that under non-food crops may fall but not necessarily in the same proportion. This is an example of limited negative correlation.

Thus, correlation is *perfect positive* if an increase (or decrease) in one variable is always followed by a corresponding and proportional increase (or decrease) in the other related variable. It is *perfect negative* if an increase (or decrease) in one factor is followed by a corresponding and proportional decrease (or increase) in the other factor. There is *no correlation* at all if values in one variable cannot be associated with values in the other variable. In between perfect positive correlation and no correlation there may be *limited degrees of positive* correlation. Similarly, in between no correlation and perfect negative correlation there may be *limited degrees of negative* correlation.

Then, we may construct a scale which begins at the top with perfectly positive correlation, passes through limited degrees of positive correlation, reaches the entire absence of correlation, and passing through limited degrees of negative correlation ends at perfectly negative correlation. Such a scale is provided by Co-efficient of Correlation.

Limits of Co-efficient of Correlation.

Coefficient of correlation is the numerical measure of the amount of correlation existing between two variables, subject and relative. That variable which is used as the standard is called the subject, and the variable which is compared with the

subject or measured in terms of the subject is called the relative. Generally, Karl Pearson's co-efficient of correlation is used. This co-efficient varies between +1 and −1. When the co-efficient reaches unity it is assumed to be perfect. Perfect positive correlation is indicated by +1, perfect negative by −1, no correlation or complete independence by o, limited degrees of positive correlation by values lying between o and +1, and limited degrees of negative correlation by values intermediate between o and −1.

Karl Pearson's Formula.

To determine the *degree of correlation* between two related variables the co-efficient of correlation devised by Karl Pearson, the great biologist and statistician, is the most satisfactory. This co-efficient is calculated by dividing the product of all the deviations of each pair of observations from their respective means by the product of the standard deviations of the two variables and the number of items. Thus, if x_1, x_2, x_3, etc., be the deviations of the values of the first variable, the subject, from the arithmetic average, and y_1, y_2, y_3, etc., be the corresponding deviations of the values of the second variable, the relative, and the summation of the products of x_1 with y_1, of x_2 with y_2, of x_3 with y_3, and so on be represented by Σxy; and, further, if the standard deviation of the subject be σ_1 and that of the relative σ_2, and n be equal to the number of pairs of observations, then the formula for measuring Karl Pearson's coefficient of correlation represented by r will be:

$$r = \frac{\Sigma xy}{n\sigma_1 \ \sigma_2}$$

When Σxy is positive, r will be positive; when Σxy is negative, r will be negative. It is the numerator which largely regulates the size of the coefficient. If positive items in one series are associated with positive items in the other series, or if negative items in one series are associated with negative items in the other series, the coefficient of correlation, r, will be positive. This means that if items larger than the arithmetic average in the subject are associated with items larger than the arithmetic average in the relative, or items smaller than the arithmetic average in the subject are associated with items

smaller than the arithmetic average in the relative, the correlation coefficient will be positive. If, however, positive items in one series are associated with negative items in the other or *vice versa*, the correlation co-efficient will be negative. When positive and negative deviations in the two series are indifferently associated, correlation will tend to zero, and will reach that limit when the negative products of deviations will be equal to the positive products of deviations, *i.e.*, when Σxy will be zero.

Pearson's Correlation Co-efficient by Direct Method.

Example 1. Required to calculate co-efficient of correlation between ages of husband and wife in a given community at a certain time, using the direct method.

Table 53. *Calculation of Karl Pearson's Coefficient of Correlation between ages of husband and wife.*

Subject X			Relative Y			
Age of husband (Years) m_1	Deviation from average (25 years). x	Square of deviation x^2	Age of wife (Years) m_2	Deviation from average (18 years). y	Square of deviation y^2	Product of deviations of husband's age and of wife's age xy
19	−6	36	14	−4	16	+24
21	−4	16	16	−2	4	+ 8
22	−3	9	15	−3	9	+ 9
23	−2	4	14	−4	16	+ 8
23	−2	4	17	−1	1	+ 2
24	−1	1	14	−4	16	+ 4
24	−1	1	17	−1	1	+ 1
25	0	0	18	0	0	0
26	+1	1	17	−1	1	− 1
26	+1	1	20	+2	4	+ 2
27	+2	4	21	+3	9	+ 6
28	+3	9	20	+2	4	+ 6
28	+3	9	22	+4	16	+12
29	+4	16	22	+4	16	+16
30	+5	25	23	+5	25	+25
$\Sigma m_1=$ 375 $a_1=$ 25		$\Sigma x^2=$ 136	$\Sigma m_2=$ 270 $a_2=$ 18		$\Sigma y^2=$ 138	$\Sigma xy=$ +122

n, or number of pairs of observations$=15$

Standard deviation is determined by the formula[1], $\sqrt{\frac{\Sigma d^2}{n}}$.

In table 53, the d's in the X series are called x's; those in Y series, y's. Accordingly, the formula for computing the standard deviation of the X series is $\sqrt{\frac{\Sigma x^2}{n}}$; for that of the Y series, $\sqrt{\frac{\Sigma y^2}{n}}$.

$$\therefore \sigma_1 = \sqrt{\frac{\Sigma x^2}{n}} = \sqrt{\frac{136}{15}} = 3.011 \text{ years.}$$

$$\text{and } \sigma_2 = \sqrt{\frac{\Sigma y^2}{n}} = \sqrt{\frac{138}{15}} = 3.033 \text{ years.}$$

$$r = \frac{\Sigma xy}{n\ \sigma_1\ \sigma_2}$$

$$= \frac{+122}{15 \times 3.011 \times 3.033}$$

$$= \frac{122}{136.99}$$

$$= +0.89$$

Alternative Method.

In the above solution, we have calculated the two standard deviations first, and then multiplied their product with the number of items, in the denominator. This procedure can be abbreviated as under:

---

[1] See Table 23, Chapter X, page 172 for computing Standard Deviation by the Direct Method.

$$r = \frac{\Sigma xy}{n\sqrt{\frac{\Sigma x^2}{n}}.\sqrt{\frac{\Sigma y^2}{n}}}$$

$$= \frac{+122}{15\times\sqrt{\frac{136}{15}}.\sqrt{\frac{138}{15}}}$$

$$= \frac{122}{15\times\frac{1}{15}\sqrt{136\times138}}$$

$$= \frac{122}{15\times\frac{1}{15}\times137}$$

$$= \frac{122}{137}$$

$$= +0.89$$

The above alternative method, in effect, means

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2.\ \Sigma y^2}}$$

This alternative method is known as the Product Moment Formula of Coefficient of Correlation. In practice, however, it will be well to proceed as we have done in the main solution of example 1.

0.89 indicates a very high degree of positive correlation, implying that in most cases the age of wife varies with that of husband.

Pearson's Co-efficient by Short-cut Method.

In the above example the averages of the ages of husbands and wives were whole numbers. Therefore, the calculation of the deviations of ages, the mean, their squaring up, and their

multiplication did not involve any trouble. If, however, the averages contain a fraction, these calculations would involve much labour, to do away with which the short-cut method may be used. In using it, any whole number may be assumed as the average, deviations from it calculated, and squared, and the standard deviations computed according to the short-cut method of computing the standard deviation. The deviations of the two series are multiplied and summated. The resulting Σxy should later be corrected by subtracting from it the product of the differences between the true means and the assumed means of the two series. Thus, if p be the true average of the products, *i.e.*, true or corrected value of Σxy, then

$$p = \frac{\Sigma xy}{n} - \left[(a_1 - x_1)\ (a_2 - x_2) \right]$$

where, a_1 stands for the true average and x_1 for the assumed average of the first series, and a_2 stands for the true average and x_2 for the assumed average of the second series, and Σxy is the summation of the products of deviations from assumed means.

Then, the co-efficient of correlation, or, $r = \frac{p}{\sigma_1\ \sigma_2}$

The above two processes may also be combined into one formula, so that without changing what the symbols have so long stood for,

$$r = \frac{\Sigma xy - n[(a_1 - x_1)\ (a_2 - x_2)]}{n\ \sigma_1\ \sigma_2}$$

Example 2. Required to calculate the co-efficient of correlation between birth-rate and death-rate of a few countries of the world during 1931, using the short-cut method.

Table 54. *Calculation of Karl Pearson's Co-efficient of Correlation between birth-rate and death-rate for a few countries of the world for 1931.*

Country.	Birth rate m_1	Deviation from assumed average (26) x	Square of deviation x^2	Death rate m_2	Deviation from assumed average (15) y	Square of deviation y^2	Product of deviations of birth rate and death rate xy
Egypt	44	+18	324	27	+12	144	+216
Canada	24	− 2	4	11	− 4	16	+ 8
U. S. A.	19	− 7	49	12	− 3	9	+ 21
India	33	+ 7	49	24	+ 9	81	+ 63
Japan	32	+ 6	36	19	+ 4	16	+ 24
Germany	16	−10	100	11	− 4	16	+ 40
France	18	− 8	64	16	+ 1	1	− 8
I. F. State	20	− 6	36	14	− 1	1	+ 6
U. K.	16	−10	100	12	− 3	9	+ 30
U. S. S. R.	40	+14	196	18	+ 3	9	+ 42
Australia	20	− 6	36	9	− 6	36	+ 36
Newzealand	18	− 8	64	8	− 7	49	+ 56
Palestine	53	+27	729	23	+ 8	64	+216
Sweden	17	−11	121	12	− 3	9	+ 33
Norway	15	− 9	81	11	− 4	16	+ 36
$n=15$	$\Sigma m=385$ $a_1=25.666$		$\Sigma x^2=1989$	$\Sigma m_2=227$ $a_2=15.133$		$\Sigma y^2=476$	$\Sigma xy=+819$

n, or number of pairs of observations$=15$.

a_1 or true arithmetic average for the first series

$$=\frac{\Sigma m_1}{n}=\frac{385}{15}=25{\cdot}666$$

Let x_1, or assumed average for the first series$=26$

a_2 or true arithmetic average for the second series

$$=\frac{\Sigma m_2}{n}=\frac{227}{15}=15{\cdot}133$$

Let x_2, or assumed average for the second series=15 Standard deviation, using the short-cut method, is determined by the formula[2], $\sqrt{\frac{\Sigma d_x^2 - n\ (a-x)^2}{n}}$. In table 54, d_x 's in the first series are called x's; those in the second series y's. Accordingly, the formula for computing standard deviation of the first series is $\sqrt{\frac{\Sigma x^2 - n\ (a_1 - x_1)^2}{n}}$ and for that of the second series, $\sqrt{\frac{\Sigma y^2 - n\ (a_2 - x_2)^2}{n}}$

$$\therefore \sigma_1 = \sqrt{\frac{1989 - 15\ (25{\cdot}666 - 26)^2}{15}} = 11{\cdot}51$$

$$\text{and } \sigma_2 = \sqrt{\frac{476 - 15\ (15{\cdot}133 - 15)^2}{15}} = 5{\cdot}631$$

$$r = \frac{\Sigma xy - n[(a_1 - x_1)\ (a_2 - x_2)]}{n\ \sigma_1\ \sigma_2}$$

$$= \frac{819 - 15(-0{\cdot}334 \times 0{\cdot}133)}{15 \times 11{\cdot}51 \times 5{\cdot}631}$$

$$= \frac{819{\cdot}67}{972{\cdot}2}$$

$$= +0{\cdot}84$$

Alternative Method.

An alternative method, which does away with the necessity of separately calculating the two standard deviations, is also available. The formula used is:

$$r = \frac{\Sigma xy - \frac{\Sigma x.\ \Sigma y}{n}}{\sqrt{\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)\left(\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right)}}$$

---

[2] See Table 25, Chapter X, page 175 for computing Standard Deviation by the Short-cut Method.

Using the above formula in the case of example 2, we have,

$$r = \frac{819 - \frac{-5 \times +2}{15}}{\sqrt{\left(1989 - \frac{(-5)^2}{15}\right)\left(476 - \frac{(2)^2}{15}\right)}}$$

$$= \frac{819{\cdot}67}{\sqrt{945358{\cdot}6}}$$

$$= \frac{819{\cdot}67}{972{\cdot}2}$$

$$= +0{\cdot}84$$

The alternative method, it will be seen, combines all the different steps taken in the original solution and leads directly to the coefficient of correlation. For a clearer grasp, and, at any rate, in all elementary work, it will, however, be better to follow the method used in the main solution of example 2.

+0·84 denotes a very high degree of positive correlation between birth-rate and death-rate of the given countries. It shows that generally the death-rate varies with the birth-rate.

Co-efficient of Correlation for Long-Time Changes.

In the above two examples the variations in the items relate to a specific time. Correlation may also be studied for historical data, that is, data stretched over a period of time. Historical data may relate to (i) Long-time changes and (ii) Short-time oscillations. In computing the co-efficient of correlation or long-time changes, the method used in example 1, or if need be, that used in example 2, shall be followed throughout, the items and deviations from the mean for the same date being paired together. In computing the co-efficient of correlation for short-time oscillations this method will be modified.

Pearson's Modified Co-efficient for Short-Time Oscillations.

It is possible that the short-time changes in two variables may be in opposite directions while the long-time changes may be in the same direction. Then, if co-efficient of correlation of such variables is computed by the method used in the fore-

going two examples, a large positive co-efficient would result which would not take any account of the opposite direction of the short-time oscillations. Correlation co-efficient computed from actual items would consequently be misleading. We should, therefore, be concerned with short-time oscillations only and rid our data of the long-time variations.[3] To do it we should discover the trend and eliminate it by computing the deviations of original items from the trend. These deviations should be multiplied together to yield Σxy. And these deviations, again, should be squared up to compute standard deviations. Thus, the modification made in the original formula is that deviations of the items are taken from the trend instead of from the arithmetic average. Example 3 demonstrates the working of this method.

Example 3. Required to compute the co-efficient of correlation of the short-time oscillations for indices of supply and price of a certain commodity in Table 55.

[3] See Chapter XIV for 'Elimination of long-time variations.'

Table 55. *Computation of co-efficient of Correlation of Short-time Oscillations between Supply and Price.*

	Supply X				Price Y				
Year	Index of Supply	5-yearly moving average of Indices	Deviations from moving average	Square of deviation	Index of Price	5-yearly moving average of Indices	Deviation from moving average	Square of deviation	Product of deviations of Supply and of Price
			x	x^2			y	y^2	xy
1920	91				117				
1921	98				97				
1922	95	94	+1	1	102				
1923	92	95	—3	9	108	106	— 4	16	— 4
1924	93	96	—3	9	105	102	+ 6	36	—18
1925	96	98	—2	4	96	98	+ 7	49	—21
1926	102	100	+2	4	77	91	+ 5	25	—10
1927	107	101	+6	36	68	85	— 8	64	—16
1928	104	102	+2	4	77	82	—14	196	—84
1929	98	103	—5	25	93	81	— 4	16	— 8
1930	100	105	—5	25	89	82	+11	121	—55
1931	108	107	+1	1	83	84	+ 5	25	—25
1932	116	110	+6	36	78	85	— 2	4	— 2
1933	114				84	85	— 7	49	—42
1934	111				93				
		$n=11$		$\Sigma x^2=$ 154		$n=11$		$\Sigma y^2=$ 601	$\Sigma xy=$ —285

[Five-yearly cycle has been assumed in the above series and decimals have been ignored in computing the moving average. Greater precision could be achieved by carrying out the decimals.]

n or number of pairs of observations$=11$, since only the years 1922 to 1932 can be used in computing the co-efficient.

$$\sigma_1=\sqrt{\frac{\Sigma x^2}{n}}=\sqrt{\frac{154}{11}}=\sqrt{14}=3.742$$

$$\sigma_2=\sqrt{\frac{\Sigma y^2}{n}}=\sqrt{\frac{601}{11}}=\sqrt{54{\cdot}6}=7{\cdot}389$$

$$r=\frac{\Sigma xy}{n\ \sigma_1\ \sigma_2}=\frac{-285}{11\times3{\cdot}742\times7{\cdot}389}=-0{\cdot}937$$

$-0{\cdot}937$ denotes a very high degree of inverse correlation between supply and price for short period, indicating that as supply increases price falls in a large majority of cases, and *vice versa*.

Co-efficient for Cyclical Fluctuations.

In example 3, co-efficient of correlation has been computed for the short-time oscillations arrived at by eliminating the long-period trend from the original items. These short-time oscillations comprise seasonal, cyclical and irregular variations. If it is intended to find out the correlation between the cyclical variations of two series, it will be necessary to reduce the given cyclical fluctuations to a comparable basis by dividing each of them by the standard deviation of the series to which they belong. This process will reduce them to a common denominator. The resulting figures are then multiplied, one with the other, in each pair. The algebraic sum of the products divided by the number of paired items, gives the co-efficient of correlation. The formula $r=\frac{\Sigma\, xy}{n\ \sigma_1\ \sigma_2}$ for computing co-efficient of correlation from the original data thus becomes $r=\frac{\Sigma xy}{n}$ for cyclical fluctuations.

In actual working, it is not the cyclical variations as such which are taken into account; cycle percentages are really used. The procedure involved in arriving at the cycle percentages is as under:—

(1) Compute the *trend* of the original series.
(2) Compute the *percentage of trend* by dividing the original series by the trend and multiplying by 100.
(3) Compute the *seasonal variation* of the series in percentage terms.
(4) Compute the *cyclical variation* by subtracting "seasonal variation percentage" from "percent of trend" and multiplying by 100.

(5) Compute the *cycle per cents* by dividing the "cyclical variation" by the "standard deviation of the series."

The "cycle per cents" of a series so calculated may be correlated with the "cycle per cents" of another series similarly computed, using the formula $r = \frac{\Sigma xy}{n}$

Example 3A. Required to compute the co-efficient of correlation of the standard deviation cycles of production and employment during 1940 to 1952. (The data chosen are arbitrary).

Table 55A. *Computation of Co-efficient of Correlation of Standard Deviation Cycles of Production and Employment.*

Year	Production x	Employment y	Product of S. D. Cycles xy
	σ	σ	
1940	0.46	−0.34	−0.1564
1941	0.75	1.20	+0.9000
1942	−1.05	0.16	+0.1680
1943	−0.48	0.37	−0.1776
1944	0.55	0.31	+0.1705
1945	0.59	−0.37	−0.2183
1946	−0.61	−0.75	+0.4576
1947	−0.23	−0.41	+0.0943
1948	1.25	1.05	+1.3125
1949	1.29	1.66	+2.1414
1950	−1.27	−0.90	+1.1430
1951	0.82	0.37	+0.3034
1952	−2.07	−2.35	+4.8645
			$\Sigma xy = 10.6669$

$$r = \frac{\Sigma xy}{n}$$

$$= \frac{10{\cdot}6669}{13}$$

$$= +0{\cdot}8205$$

It will be useful to remember that correlation of time-series is best measured by using series from which both the trend and the seasonal variations have been eliminated. The reason for this is simple. In time series the relationship of one item to the other is due, among other things, to long-time and short-time influences of different types. Accordingly, if it is intended to measure the degree of correlation between time series, it is useless to correlate them until, so far as possible, the different types of fluctuations have been isolated. Instead, if the co-efficient of correlation is computed for the original series and a value different from zero arrived at, it would be difficult to say what contributed most largely to the result—similar or diverse trends, seasonal variations, cyclical fluctuations or irregular movements. Therefore, before correlating time-series, the trend and the seasonal variations should be eliminated from them. The cycle percentages are certainly less "ordered in time", to borrow Pearsons' phrase, than are the original items of series. Business and economic data have been found to be much more alike in their cyclical relations, than in all of their fluctuations. That is, whereas their secular trends and seasonal fluctuations are peculiar to themselves, their cyclical variations are the results of the general economic and business conditions.

Correlation Co-efficient of Grouped Series.

In the first three examples, the given series relate to quantitative individual observations. Correlation of grouped series can also be similarly studied. We may measure an adequate number of pairs of values for each member and find what values are associated together and how often the same values are repeated. When this is done, we can group our data into a table of double entry, or contingency table. Suppose we find that in two class-tests—one in Economics and the other in Geography—at which 60 boys were examined the following were the results:—

Table 56. *Frequency distribution of Marks in Economics & Geography.*

X Economics (Max. Marks 50)		Y Geography (Max. Marks 50)	
Marks obtained	Number of boys	Marks obtained	Number of boys
5—15	5	0—10	2
15—25	18	10—20	15
25—35	27	20—30	20
35—45	10	30—40	15
		40—50	8
	60		60

If we desire to study the relationship between the knowledge of Economics and that of Geography with the help of the above two series, we would need some more information: we should know what values of the two series are associated together and how frequently the same values are repeated. Suppose we find that one boy who got marks varying between 5—15 in Economics also got marks varying between 0—10 in Geography, that three boys who got marks varying between 5—15 in Economics also got marks varying between 10—20 in Geography, and so on, we can prepare a table of double entry as follows:—

Table 57. *Correlation Table for marks in Economics and Geography.*

Y Marks in Geography (max. marks 50)	X Marks in Economics (max. marks. 50)				Total
	5—15	15—25	25—35	35—45	f_y
0—10	1	1			2
10—20	3	6	5	1	15
20—30	1	8	9	2	20
30—40		3	9	3	15
40—50			4	4	8
Total f_x	5	18	27	10	60

Example 4. Required to compute correlation co-efficient from the data given in table 57.

[To compute the co-efficient the formula used in example 1 where deviations were calculated from the true mean, or the formula used in exercise 2 where deviations were calculated from the assumed mean, may be used in this example too. The latter procedure saves much labour and, therefore, it will be adopted in the given case.]

In tables 58 and 59 below we calculate the standard deviations of the X and the Y series relating to marks in Economics and Geography, respectively. Let the assumed averages, x_1 and x_2, for the X and the Y series be respectively 30 and 25 marks.

Table 58. *Calculation of Standard Deviation of Y series.*

Marks group	Mid-value	Frequency	Product of mid-value & frequency	Deviation from assumed average (30)	Square of deviation	Product of frequency & square of deviation
	m	f	mf	d_x	$d^2{}_x$	$fd^2{}_x$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
5—15	10	5	50	—20	400	2000
15—25	20	18	360	—10	100	1800
25—35	30	27	810	0	0	0
35—45	40	10	400	+10	100	1000
		$n=60$	$\Sigma m=$ 1620			$\Sigma d^2{}_x=$ 4800

x_1 or Assumed Average$=30$ marks.

a_1 or True Average$=\dfrac{\Sigma m}{n}=\dfrac{1620}{60}=27$ marks.

$$\sigma_1=\sqrt{\frac{\Sigma d^2{}_x-n\ (a_1-x_1)^2}{n}}$$

$$=\sqrt{\frac{4800-60\ (27-30)^2}{60}}$$

$$=\sqrt{71}$$

$$=8{\cdot}426 \text{ marks.}$$

Table 59. *Calculation of Standard Deviation of Y series.*

Marks group	Mid-alue	Frequency	Product of mid-value & frequency	Deviation from assumed mean (25)	Square of deviation	Product of frequency & square of deviation
	m	f	mf	d_y	d^2	fd^2_y
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0—10	5	2	10	−20	400	800
10—20	15	15	225	−10	100	1500
20—30	25	20	500	0	0	0
30—40	35	15	525	+10	100	1500
40—50	45	8	360	+20	400	3200
		$n=60$	$\Sigma m=$ 1620			$\Sigma d^2_y=$ 7000

x_2 or Assumed Average=25 marks.

a_2 or True Average$=\frac{\Sigma m}{n}=\frac{1020}{60}=27$ marks.

$$\sigma_2=\sqrt{\Sigma d^2{}_y-n\ (a_2-x_2)^2}$$

$$=\sqrt{\frac{7000-60\ (27-25)^2}{60}}$$

$$=\sqrt{112{\cdot}66}$$

$$=10{\cdot}614 \text{ marks.}$$

Now, the value of Σxy remains to be determined. Table 60 shows the method of determining it. d_x, deviations from the assumed mean in X series shown in column (5), table 58, and d_y, deviations from the assumed mean in Y series shown in column (5), table 59, are taken to table 60. Related pairs of deviations are first multiplied and the product put down in

Table 60. *Calculation of Summation of Products of Deviations.* $(\Sigma\, xy)$

Row No.	Column No		1	2	3	4	Product of d_x and d_y and frequency (fxy)
	Marks→X		5—15	15—25	25—35	35—45	
	↓ Y	d_x→ d_y↓	—20	—10	0	+10	
1	0—10	—20	400 1 400	200 1 200	—	—	600
2	10—20	—10	200 3 600	100 6 600	0 5 0	—100 1 —100	1100
3	20—30	0	0 1 0	0 8 0	0 9 0	0 2 0	0
4	30—40	+10	—	—100 3 —300	0 9 0	100 3 300	0
5	40—50	+20	—	—	0 4 0	200 4 80 0	800
	Product of d_x and d_y and frequency (fxy)		1000	500	0	1000	Σxy= 2500

the left-hand top corner of their respective cells. Thus,—20, d_x, is multiplied with—20, d_y, and the product placed in the left corner of the cell formed by the 1st row and the 1st column;—20, d_x, and—10, d_y, are multiplied and the product placed in the left corner of the cell formed by the 2nd row and the 1st column;—10, d_x, and +10, d_y, are multiplied and the product,—100, placed in the left corner of the cell formed by 4th row and 2nd column; and so on.

These products of d_x and d_y are multiplied by their respective frequencies placed in the centre of their respective cells. The final products are placed in the right-hand bottom corner of their respective cells. These final products, when summated give the Σxy. This summation refers to the assumed averages and will, therefore, as in example 2, be corrected by subtracting from it n times the product of the difference between true and assumed means of the two series. Thus,

$$r = \frac{\Sigma xy - n[(a_1 - x_1)\ (a_2 - x_2)]}{n\sigma_1\ \sigma_2}$$

$$= \frac{2500 - 60[(27-30)\ (27-25)]}{60 \times 10.614 \times 8.426}$$

$$= \frac{2860}{5366.01}$$

$$= +\ 0.533$$

+0.533 indicates a moderately high degree of positive correlation between Economics and Geography.

Alternative Method for S.D. of Grouped Series.

There is an alternative and easier method of calculating standard deviation in a grouped series, if the magnitude of its class-intervals is uniform. In this alternative method, the mid-points of the class-intervals need not be found out: they may be taken as 1, 2, 3, 4, 5, etc. and deviations may be computed

from any of these assumed mid-points or from their average. The remaining steps are the same as shown in Table 58 or 59 except that in the formula provision should be made for multiplication by the magnitude of the class-interval, so that the resulting standard deviation would turn out to be the same as computed by the method shown in Tables 58 and 59. The formula for this alternative method may be written thus:—

$$\sigma = \sqrt{\frac{\Sigma d^2{}_x}{n} - \left(\frac{\Sigma d_x}{n}\right)^2} \times i$$

where d_x stands for deviation from a mid-point. n for number of items and i for the magnitude of class-intervals.

Let us apply to the data given in Table 59 the above alternative method.

Table 59A. *Calculation of Standard Deviation.*

Marks group	Assumed mid-point m	Frequency f	Deviation from 3 d_x	Product of frequency and deviation fd_x	Square of deviation $d^2{}_x$	Product of frequency and square of deviation $fd^2{}_x$
0—10	1	2	—2	—4	4	8
10—20	2	15	—1	—15	1	15
20—30	3	20	0	0	0	0
30—40	4	15	+1	+15	1	15
40—50	5	8	+2	+16	4	32
		$n=60$		$\Sigma d_x = +12$		$\Sigma d^2{}_x = 70$

$$\sigma = \sqrt{\frac{\Sigma d^2{}_x}{n} - \left(\frac{\Sigma d_x}{n}\right)^2} \times i$$

$$= \sqrt{\frac{70}{60} - \left(\frac{12}{60}\right)^2} \times 10$$

$$= \sqrt{1.1266} \times 10$$

$$= 10.614 \text{ marks.}$$

So, we get the same value for the standard deviation as in Table 59.

Alternative Method for Co-efficient of Grouped Series.

The principle underlying the above alternative method of computing the standard deviation of a grouped series may well

Table 60A. *Coefficient of Correlation between marks in Economics and Geography.*

X → Y ↓	5—15	15—25	25—35	35—45	f_y xy	d_y	fd_y	d^2	fd_y^2
0—10	1 +2	1 0			2 +2	—2	—4	4	8
10—20	3 +3	6 0	5 —5	1 —2	15 —4	—1	—15	1	15
20—30	1 0	8 0	9 0	2 0	20 0	0	0	0	0
30—40		3 0	9 +9	3 +6	15 +15	+1	+15	1	15
40—50			4 +8	4 +16	8 +24	+2	+16	4	32
f_x xy	5 +5	18 0	27 +12	10 +20	60 +37		+12		70
d_x	—1	0	+1	+ 2					
fd_x	—5	0	+27	+20	+42				
d_x^2	1	0	1	4					
fd_x^2	5	0	27	40	72				

be extended to the computation of coefficient of correlation of a grouped series. The standard deviations may be calculated in the manner shown in Table 59A and Σxy in the manner shown in Table 60, but they need not be multiplied by the magnitude of class-intervals inasmuch as these magnitudes are common to all. Further, instead of resorting to three tables (like Tables 58, 59 and 60 in example 4) the entire working may be compressed into one table, as in Table 60A, which gives the solution of example 4 by the alternative method.

We get the following values from Table 60A:

$$\Sigma xy = +37$$

$$\Sigma fd_x \text{ or } \Sigma x = +42$$

$$\Sigma fd_y \text{ or } \Sigma y = +12$$

$$\Sigma fd^2_x \text{ or } \Sigma x^2 = 72$$

$$\Sigma fd^2_y \text{ or } \Sigma y^2 = 70$$

$$n = 60$$

Now,

$$r = \frac{\Sigma xy - \dfrac{\Sigma x . \Sigma y}{n}}{\sqrt{\left(\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}\right)\left(\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}\right)}}$$

$$= \frac{37 - \dfrac{42 \times 12}{60}}{\sqrt{\left(72 - \dfrac{(42)^2}{60}\right)\left(70 - \dfrac{(12)^2}{60}\right)}}$$

$$= \frac{37 - 8.4}{\sqrt{\dfrac{213}{5} \times \dfrac{338}{5}}} = \frac{28.6}{\sqrt{2879.76}}$$

$$= \frac{28.6}{53.66}$$

$$= +0.533$$

Thus, by the alternative method too we have arrived at the same value of r as by the detailed method, and the ignoring of the magnitude of class-intervals has had no effect on the answer.

In examples 1, 2 and 4 correlation is positive but not perfect. A simple example of perfect positive correlation is the following:—

Number of persons:	1, 2, 3, 4, 5, 6, 7, 8.
Number of eyes:	2, 4, 6, 8, 10, 12, 14, 16.

Correlation co-efficient in the above series will be + 1.

Assumptions of Pearsonian Correlation.

Karl Pearson's co-efficient of correlation between two variables is based on three assumptions:

(1) *In each of the series correlated a large variety of independent causes are operating so as to produce normal distribution.* Such causes, for example, are variations in climate, nourishment, physical training, environment. The series resulting from the effect of such independent contributory causation would conform to normal curve of error.

(2) *The forces so operating are related in a causal way.* If the forces are independent of each other, if no cause producing the forces is common to both series, there would be no correlation. If the height of a house remained unaltered while that of a growing child increased, there would be no correlation between them, since the causes affecting one variable would not be found to affect the other. In this case, the sizes in one series could not be said to be associated with the sizes in the other.

(3) *The relationship between the two series is linear.* If straight line relationship does not hold, the coefficient may be very small but will never be too large.

Characteristics of Pearsonian Co-efficient.

Karl Pearson's co-efficient of correlation is zero when independence between two variables is complete and is unity when there is perfect correlation, *i.e.*, when connection between the variables is rigid. It always varies between + 1 and − 1

and is a sensitive measure of the amount of correlation. It is based on all the observations of the given variables and is independent of the units in which the variables are measured.

Probable Error of the Co-efficient.

Probable error is a measure which added to or subtracted from a most probable measurement gives the limits within which it is probable that an item of the same nature, if selected *at random*, will fall.

Co-efficient of correlation also has a probable error. It is that amount which when added to and subtracted from the average correlation co-efficient gives amounts within which the probability is that a co-efficient of correlation from series selected *at random* from the same universe will fall.

The formula for the probable error of Karl Pearson's co-efficient of correlation is

$$0.6745 \frac{1-r^2}{\sqrt{n}}$$

where, r is the co-efficient of correlation and n the number of items paired.

The probable error of the co-efficient of correlation, $+0.89$, between the ages of husband and wife computed in example 1, will be

$$0.6745 \frac{1-(0.89)^2}{\sqrt{15}} = 0.036.$$

The co-efficient of correlation for the example under consideration should, therefore, be written as

$$r = +0.89 \pm 0.036.$$

We can now say that the co-efficient of correlation in the example actually lies between 0.854 and 0.926, and that another co-efficient computed from series chosen *at random* from the

universe from which the given series was selected, would fall within this range.

Interpretation of Co-efficient of Correlation.

As already pointed out, the co-efficient of correlation for a given pair of variables must lie between $+1$ (perfect direct correlation) and -1 (perfect inverse correlation). When the co-efficient is zero, it indicates complete absence of correlation. So, when the co-efficient is 0.95 or over, the degree of correlation is obviously high.

It may be asked whether correlation between a given pair of variables, as indicated by the co-efficient, is 'significant.' This can be answered in the light of the following four empirical rules based on the use of probable error of the co-efficient:

(1) If r is less than the probable error, there is no evidence of correlation.

(2) If r is more than six times the probable error, correlation is significant, that is, its existence is a practical certainty.

(3) If the probable error is relatively small, correlation should not be considered at all marked when r is less than 0.30.

(4) If the probable error is relatively small, correlation is decidedly existing when r is above 0.50.

In example 1, r is nearly 25 times the probable error. Correlation is, therefore, significant.

It may be noted that the probable error at times leads to wrong results, unless r is small and n is large. In order that the formulæ for co-efficient of correlation and probable error may yield satisfactory results, n should be considerably large.

The above four rules must be kept in view while interpreting the correlation co-efficient. When a correlation co-efficient, it may be added, is found to be significant, it should not be implied to mean more than what it does. For instance, in example 1, correlation between ages of husbands and wives is

strong positive. It simply shows a connection between the two age series and does not necessarily mean that *every* young husband has young wife. Correlation is true on the average. A particular old man may have a young wife or two wives, one young and the other old. Again, if supply and price in example 3 are negatively correlated, it does not mean that increase in supply is, or is the *only*, cause of fall in price. Similarly, if marks in Geography and Economics are positively correlated, it does not imply that the two subjects are necessarily related as cause and effect. Knowledge of one subject may be helpful in the other, but the correlation may also be due to some third factor, *e.g.*, adequate teaching in both the subjects. *So, a direct cause and effect relationship is not always and in all cases established by the fact that two series are correlated.*

Correlation Co-efficient by Method of Rank Differences.

Karl Pearson's method is sometimes applied to the rankings of the data. This method is suitable for use with irregular data where extreme items are erratic and perhaps not quite accurate. It is applied to series of individual observations rather than frequency distributions.

The method implies substituting for the given quantities, their rank. In each series the item with the largest size is ranked 1, next largest 2, and so on. If ties occur, they may be assigned the average of the ranks they would have received if they had differed slightly. Thus, if two items are tied for the third rank, each may be ranked $(3+4)\div 2$ or 3.5, and the next item in order of size would be ranked fifth. This arrangement is not entirely valid; the resulting coefficient will be less reliable, if the ties are numerous. After the data have been ranked from the largest to the smallest in each series respectively, the differences in ranks are taken. The formula for calculating the coefficient applied to these differences is an algebraic simplication of the usual formula $r=\frac{\Sigma xy}{n\sigma_1\sigma_2}$ as applied to rankings. The formula is:

$$r_r=1-6\,\Sigma\, d^2\div[n(n^2-1)]$$

where r_r is the coefficient of correlation by the method of rank differences, d is the difference in ranks of each pair of correlatives and n is the number of pairs of observations.

Example 5. Required to calculate the coefficient of correlation from the data given below, by the method of rank differences:

$X =$ 78, 89, 97, 69, 59, 79, 68, 57.
$Y =$ 125, 137, 156, 112, 107, 136, 123, 108.

Table 60B. *Computation of Coefficient of Correlation by the method of rank differences.*

X	Rank	Y	Rank	Rank Differences d	Square of Rank Differences d^2
78	4	125	4	0	0
89	2	137	2	0	0
97	1	156	1	0	0
69	5	112	6	−1	1
59	7	107	8	−1	1
79	3	136	3	0	0
68	6	123	5	+1	1
57	8	108	7	+1	1
				$\Sigma d = 0$	$\Sigma d^2 = 4$

$$r_r = 1 - \frac{6 \Sigma d^2}{n (n^2 - 1)}$$
$$= 1 - \frac{24}{504}$$
$$= +0.95$$

If Pearson's useful formula $r = \frac{\Sigma xy}{n \sigma_1 \sigma_2}$ is applied to the data in the above example, then too the co-efficient of correlation will be $+0.95$.

Co-efficient of Concurrent Deviations.

A much simpler though crude method of correlation is the method of concurrent deviations. This method is sometimes useful in a preliminary investigation. It gives a measure of association in the direction of change only. In example 3, we have used the modified method of measuring coefficient of correlation for short-time oscillations. Co-efficient of concurrent deviations provides a much simpler method for the same study and gives satisfactory results in most cases. This method, however, is not suitable for dealing with long-time changes, since it does not take account of the general trend.

If, in comparing two historical variables relating to short-time oscillations it is found that the two curves move in the same direction at the same time—that is, if the deviations are concurrent—there is a marked evidence of direct or positive correlation between the short-time fluctuations. But, if the curves are steadily moving in opposite directions—that is, if the deviations are divergent—there is an evidence of inverse or negative correlation. To compute the co-efficient of correlation in such cases, we take into consideration not the deviations from the arithmetic means nor those from the moving averages but simply from the measurement of the preceding date recorded. Secondly, we take into consideration not the size of the deviation but only its direction. The following empirical formula is used for the purpose of computing the co-efficient. This formula has the same characteristics as that of Karl Pearson, *viz.*, $+1$ denotes perfect positive correlation,-1 indicates perfect negative correlation and o shows absence of correlation.

If r=the co-efficient of correlation,
n=the number of pairs of observations,
c=the number of concurrent deviations, then

$$r = \pm\sqrt{\pm\left(\frac{2c-n}{n}\right)}$$

The use of signs should be carefully understood. If the quantity $\frac{2c-n}{n}$ is negative, a minus sign is placed before it and also before the radical so that the square root can be taken and the resulting co-efficient may retain the same sign as that of the original quantity.

Example 6. Required to calculate the co-efficient of correlation from the data given in table 55, page 416, by the method of concurrent deviations.

Table 61. *Computation of Correlation of Short-time Fluctuations of Supply and Price by means of Concurrent Deviations.*

Year	Supply X		Price Y		Product
	Index of Supply	Deviation from preceding year x	Index of Price	Deviation from preceding year y	xy
1920	91		117		
1921	98	+	97	—	—
1922	95	—	102	+	—
1923	92	—	108	+	—
1924	93	+	105	—	—
1925	96	+	96	—	—
1926	102	+	77	—	—
1927	107	+	68	—	—
1928	104	—	77	+	—
1929	98	—	93	+	—
1930	100	+	89	—	—
1931	108	+	83	—	—
1932	116	+	78	—	—
1933	114	—	84	+	—
1934	111	—	93	+	—

$n=14$. since only the years 1921 to 1934 can be used in computing the co-efficient.

$c=o$, since there are no pairs having like signs.

$$r=\pm\sqrt{\pm\left(\frac{2c-n}{n}\right)}$$

$$=\pm\sqrt{\pm\left(\frac{0-14}{14}\right)}$$

$$=\pm\sqrt{-(-1)}$$

$$=-\sqrt{1}$$

$$=-1.$$

Therefore, there is perfect inverse correlation.

CORRELATION BY GRAPHIC METHOD

Correlation Graph.

One way of detecting the negative or positive character of correlation is to draw correlation graphs and read the direction of the curves. This method is illustrated in Figure 30, below, in which monthly figures relating to volume and value of exports of rice by sea from undivided India given in table 62 are plotted.

Table 62. *Foreign Sea-borne Trade—Exports (Value and Volume) of Rice (not in husk)* in 1941-42.

Month	Volume	Value
	(000) Tons.	(00,000) Rs.
April	22	32
May	29	45
June	22	32
July	19	29
August	27	44
September ..	43	69
October	24	40
November	18	29
December	20	31
January	23	37
February	32	53
March	26	43
Average	25.4	140.3

In drawing correlation graphs the choice of scales and base line should be so made that if lines representing averages of the two series are drawn parallel to the base they would be as close to each other as possible. There is no objection to taking a false base line if it is required for bringing the two average lines nearer each other. By drawing the curves on such base line and scale their fluctuations would be thrown into proper relief.

Correlation graph representing volume and value of rice exported from India.

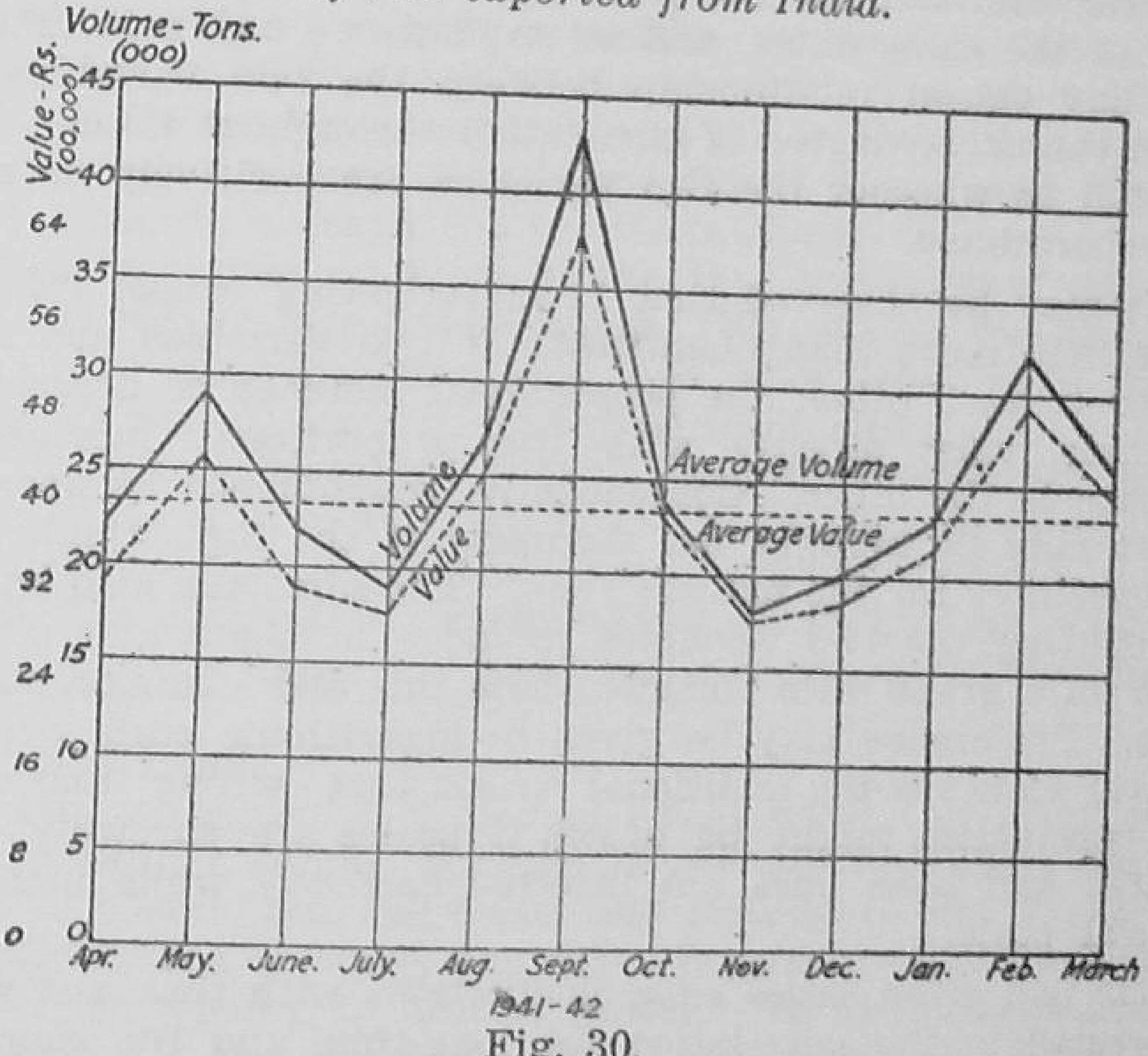


Fig. 30.

In Figure 30 the lines showing average volume and average value are close to each other. They may have been brought still closer or made to overlap each other. But, thereby the two curves, the nature of their series being such, may also have overlapped each other, at least for a large part, and their graphic display would have been spoilt. The curves, as they are drawn, serve their purpose alright. They run remarkably

parallel to each other throughout their upward and downward journey. They, therefore, indicate positive correlation between volume and value of exports of rice during the several months of 1941-42. Similarly, with another correlation graph we may have found that both the curves steadily ran in opposite directions, yielding negative correlation between the series.

But correlation graphs are not capable of doing anything more than suggesting the fact of a possible relationship between two variables. We can certainly note from them whether fluctuations agree throughout their courses, whether both of them rise and fall together, whether maxima and minima occur at the same dates, and so on; but, we can neither establish any causal relationship between the two variables nor obtain the exact degree of correlation throughout them. They only tell us whether the two variables are positively or negatively correlated.

It may be observed that in investigating causal relations ratios help more than quantities. If two variables are really related to each other, the proportional increase or decrease in one may vary directly with the proportional increase or decrease in the other. Consequently, resemblance between two curves may be brought out distinctly if the logarithmic scale is substituted for the natural scale. The same can also be done by reducing the two variables to index numbers and plotting them on a graph with common base line and common vertical scale. Preference may be given to logarithmic scale over the natural scale for the additional reason that wrong and deceptive conclusions might be drawn if scales are shrewdly manipulated and base lines not inserted correctly while using the natural scale.

If two given series show fluctuations with time and we are interested in the correlation for long-time and for short-time changes a different method of comparison will be followed.

Graphic Correlation of Time Changes.

In example 3 we have used a modification of Pearson's method for computing the correlation co-efficient. The reason for the modification is that if we desire to discover the relationship of short-time oscillations we should rid our data of the

long-time variations. If the annual index numbers of supply and price in table 55 are plotted on a graph and their correlation is studied, it would be found that it is negative, the two curves moving in opposite directions. If long-period changes are compared by plotting the 5-yearly averages given in columns 3 and 7, it will be seen that as supply shows a rising trend from 1920 onwards the price shows a downward trend from 1920 upto 1928; but, after 1928 even with an upward trend of supply the trend of the price is also upward, which fact may be due to increase in population, prosperity or change in demand. A comparison of the long-term changes after 1928 would, therefore, suggest a positive correlation, while that of those before 1928, a negative correlation. But when the long-time variations are eliminated and we are left with short-time oscillations, as given in columns 4 and 8, we may plot these deviations on a graph (as we did in Figure 29) and observe a marked relationship between the two curves, now completely unobscured by long-time changes. We would see that when one curve rises the other falls and *vice versa*. The fact of negative correlation would thus be made clear.

It follows, therefore, that to study the correlation for long-time changes in a time series we should plot the moving averages of the two series and compare their directions, while to study the correlation for short-time oscillations we should plot the deviations from the trend and observe their movement.[4]

Lag and Lead.

Not infrequently it is found that while two given historigrams exhibit similar movements thereby indicating the existence of correlation, the crests and troughs of the waves in the graphs do not coincide, one curve lagging behind the other or one curve tending to lead or precede the other. For instance, increment in the supply of money is accompanied by a rise in the general level of prices, which in its turn leads to a rise in the cost of living, but it takes time for rise in the cost of living to result from increment in the supply of money, so that money supply curve will precede and cost of living curve will lag.

---

[4] See in this connection 'Comparison of time-changes', Chapter XIV, pages 393-394.

When lag is known to exist, it will be well to allow for it by lagging the "effect" curve so that the curve may be made comparable with the "cause" curve. An easy way of determining the period of lag is to plot the two series on one graph—*i.e.*, draw correlation graph—and read the time-distance between the peaks of the two curves. Suppose cost of living index shows a rise three months after the rise in the index of supply of money, then the period of lag is three months. To allow for it, the cost of living curve should be lagged by three months. Thus, in computing the co-efficient of correlation, the index of money for January should be paired with the index of cost of living for April, index of money for February with index of cost of living for May, and so on. The underlying principle is that the dates coupled together should be not identical but so different that the resulting correlation will be the highest.

Scatter Diagram.

The data relating to volume and value of exports of rice from undivided India given in Table 62 are presented by means of scatter diagram in Figure 30A. First the co-ordinates are plotted and we get as many points on the graph as the number of pairs of items. A *line of best fit* is then drawn freehand so as to lie as closely as possible to all the plotted points. A good method would be to stretch a piece of thread and find out the best position for the line to take.

If there is high direct correlation between two variables, the line of best fit slopes upwards from left to right. This is the case with Fig. 30A. If there is high inverse correlation, the line slopes downwards from left to right. Data given in Table 61 would reveal this tendency if presented on a scatter diagram. If there is little or no correlation, the plotted points are scattered widely over the whole graph so that no line of best fit is possible.

An advantage of the scatter diagram is that the value in one series corresponding to a given value in the other series can be obtained from the line of best fit. It will be seen from Fig. 30A that if the exports of rice were 30,000 tons, the value thereof would probably be Rs. 49,00,000.

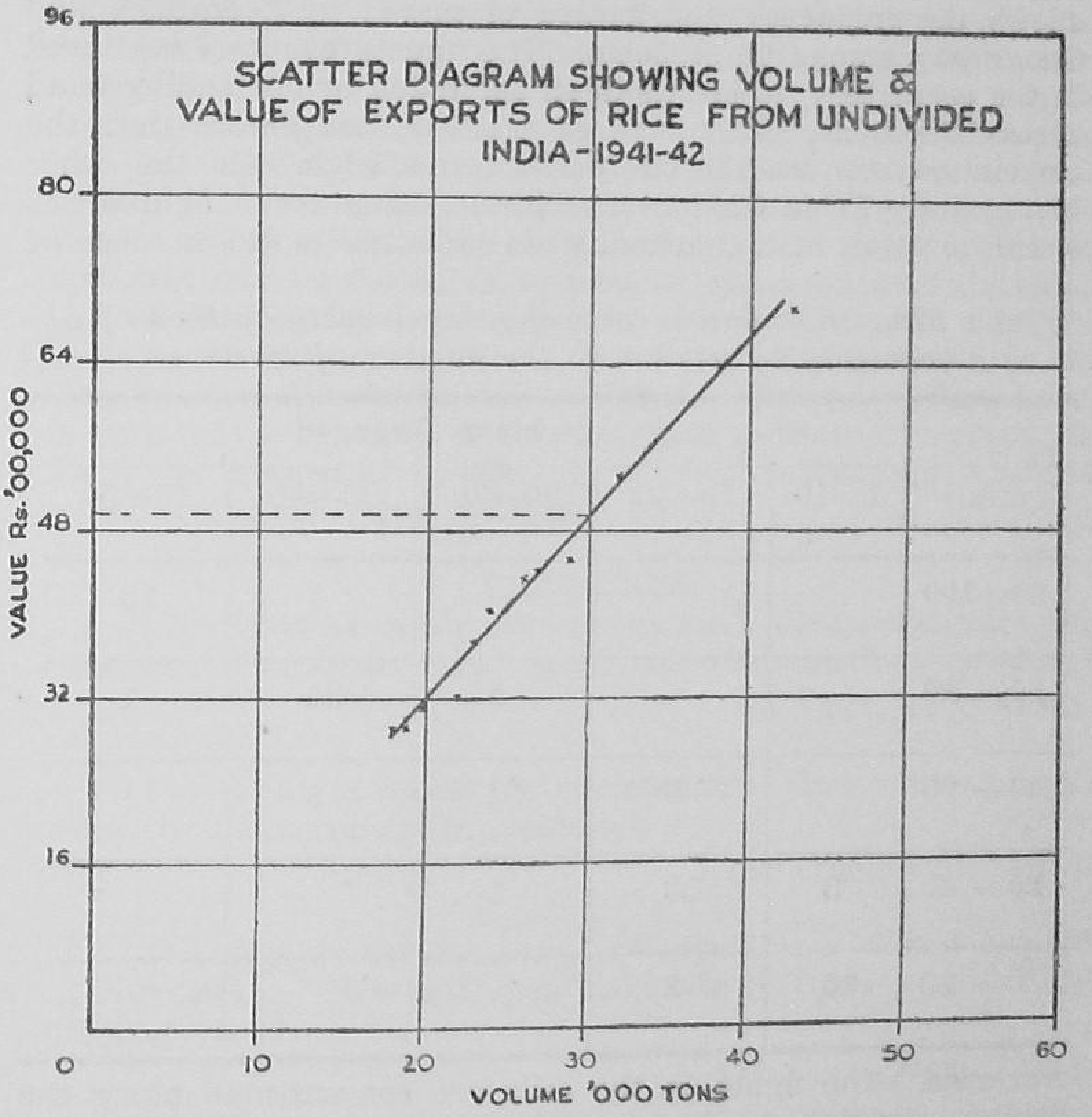

SCATTER DIAGRAM SHOWING VOLUME & VALUE OF EXPORTS OF RICE FROM UNDIVIDED INDIA - 1941-42
VALUE Rs. '00,000
96
80
64
48
32
16
0
10
20
30
40
50
60
VOLUME '000 TONS

Fig. 30A.

CORRELATION TABLE

One method of studying correlation in a grouped frequency distribution is to draw a table of double entry and observe the scatteredness of frequencies in the cells of the table. Table 57 shows the frequency distribution of marks in Economics and Geography among 60 students. The figures are well scattered in the cells; they do not lie in an ascending or descending band across the table. This type of scatteredness shows that the correlation between the two series is not high. On the other hand, study Table 62A, which gives imaginary figures of weekly outturn of a commodity in relation to the number of workmen. The figure in the cells are concentrated along the diagonal, leading to the conclusion that the correlation is high and positive.

Table 62A. *Correlation table showing Weekly outturn of a Commodity in relation to the number of workmen.*

Outturn (mds.)	Workmen Engaged				
	1—10	10—20	20—30	30—40	40—50
80—100				3	15
60— 80			6	34	1
40— 60		5	51	6	
20— 40	3	33	1		
0— 20	16	3			

It is interesting to note that the arrangement of figures in a correlation table is very much akin to the dots on a scatter diagram. The table has the added advantage that it shows the frequencies falling under each class-interval.

Methods of Studying Correlation compared.

Three methods of studying correlation have been discussed in the foregoing pages. Both are graphic method and the correlation table indicate the fact of correlation; they do not measure it. Karl Pearson's co-efficient of correlation does both; it indicates the fact of correlation and at the same time measures it. The method of correlation table is applicable only to a grouped frequency distribution, where the graphic method is not very useful. The method of co-efficient of correlation is applicable to series of individual observations, grouped frequency distributions, time-series and all. The graphic method no doubt enables the superimposing of one series over the other to find out the appropriate period of lag, which will give the highest degree of correlation. But the graphic method does not measure the correlation for different timings. This again is done only by the use of the numerical measure of correlation. Thus the method of co-efficient of correlation is superior to either of the other two methods.

EXERCISES

(1) What is meant by correlation? Does it signify the existence of cause and effect relationship between two variables?

(M. Com., Raj., 1952).

(2) What is meant by correlation? Give the general rules for interpreting its coefficient.

..(M. Com., Alld., 1944).

(3) Explain the meaning and significance of the concept of correlation. How will you calculate it from a statistical point of view?

(M. Com., Agra, 1945).

(4) What is correlation? Explain how you will use the following methods in determining correlation:

(i) Graph, (ii) Correlation Table, (iii) Karl Pearson's coefficient of correlation.

(B. Com., Agra, 1940).

(5) Given:—
Number of pairs of observations of X and Y series=15.

X Series :	Arithmetic Average	=	25.0
	Standard Deviation	=	3.01
Y Series :	Arithmetic Average	=	18.0
	Standard Deviation	=	3.03

Summation of products of corresponding deviations of X & Y series = +122.0.

Calculate (a) the coefficient of correlation between X and Y series, (b) the probable error of the coefficient.

(B. Com., Alld., 1948).

(6) Calculate the coefficient of correlation between production of wheat and its price from the following table:—

Year	Production (millions of bushels)	Price per bushel (Cents).
1890	449	84
1895	542	51
1900	599	62
1905	706	75
1910	625	91
1915	1,008	96
1920	843	183
1925	669	144
1930	886	67
1935	626	83

(M.A., Patna, 1941).

(7) The following table gives the annual birth rates and death rates in U.S.A. during the period 1931 to 1945:—

Year	Birth rate	Death rate
1931	18.0	11.1
1932	17.4	10.9
1933	16.6	10.7
1934	17.2	11.1
1935	16.9	10.9
1936	16.7	11.6
1937	17.1	11.3
1938	17.6	10.6
1939	17.3	10.6
1940	17.9	10.7
1941	18.9	10.5
1942	20.9	10.4
1943	21.5	10.9
1944	20.2	10.6
1945	19.6	10.6

Calculate the coefficient of correlation between birth rates and death rates in U. S. A.

(M. Com., Raj., 1951).

(8) Calculate Karl Pearson's Coefficient of Correlation from the following data:—

Year	Index of rainfall	Index of crop production
1930	148	137
1931	145	137
1932	141	133
1933	116	108
1934	96	95
1935	91	99
1936	87	97
1937	89	96
1938	91	99
1939	91	102
1940	102	108
1941	95	104
1942	108	108
1943	120	116
1944	139	120

(M.A., Punjab, Apr., 1952).

(9) Calculate the coefficient of correlation between the national income and newspaper circulation from the data given below:—

Year	Newspaper circulation (millions of copies)	National income (millions of dollars)
1930	39.6	68.9
1931	38.8	54.5
1932	36.4	40.0
1933	35.2	42.3
1934	36.7	49.5
1935	38.2	55.7
1936	40.3	64.9
1937	41.4	71.5
1938	39.6	64.2
1939	39.7	70.8
1940	41.1	77.5

(M.A., Alld., 1951).

(10) From the following table compute the coefficient of correlation between savings bank deposits and strikes and lock-outs over a period of seven years.

Savings Bank deposits in lakhs of rupees	51	54	55	59	65	60	70
Strikes and lockouts	38	44	33	36	33	23	10

(B. Com., Hons., Andhra, 1944).

(11) Calculate the coefficient of correlation between cotton cloth manufactures and woollen manufactures from the following data for July 1951—June 1952.—

Index of Industrial Production India (Base' 1946=100)

Month	Cloth manufactures	Woollen manufactures
1951-52		
July	105	64
August	108	63
September	106	66
October	100	62
November	104	69
December	105	61
January	104	57
February	111	55
March	104	47
April	114	45
May	118	45
June	117	50

(B. Com., Alld., 1953).

(12) Calculate the coefficient of correlation between the values of X & Y given below:—

X	Y
78	125
89	137
97	156
69	112
59	107
79	136
68	123
57	108

(You may use 69 as working mean for X and 112 as that for Y).

(M.A., Delhi, 1953).

(13) The index numbers of prices of all commodities in Bombay and in Calcutta were as under:—

Month	Index number of Commodity prices in Calcutta	Index number of Commodity prices in Bombay
May 1942	169	204
June 1942	182	222
July 1942	182	225
August 1942	192	228
September 1942	198	229
October 1942	209	233
November 1942	227	249
December 1942	238	266
January 1943	250	255
February 1943	253	255

Do you think prices in Bombay and in Calcutta are correlated?

(M.A., Agra, 1944).

(14) The following table gives the value of exports of raw cotton from India and the value of the imports of manufactured cotton goods into India during the years 1913-14 to 1931-32:—

Year		Exports of Raw Cotton	Imports of manufactured Cotton Goods
1913-14	..	42	56
1917-18	..	44	49
1919-20	..	58	53
1921-22	..	55	58
1923-24	..	89	65
1929-30	..	98	76
1931-32	..	66	58

Calculate the co-efficient of correlation between the value of the exports of raw cotton and the value of the imports of cotton manufactured goods.

(M.A., Cal., 1937).
(B. Com., Nag., 1945).

(15) The following table gives the wholesale price index numbers for Calcutta and Karachi for the period 1927-1941:—

Year	Calcutta Index Numbers (Base: July, 1914)	Karachi Index Numbers (Base: July, 1914)
1927	148	137
1928	145	137
1929	141	133
1930	116	108
1931	96	95
1932	91	99
1933	87	97
1934	89	96
1935	91	99
1936	91	102
1937	102	108
1938	95	104
1939	108	108
1940	120	116
1941	139	120

(*a*) Calculate the co-efficient of correlation between the above two series, and state what it indicates.

(*b*) Point out whether the Calcutta indices are more variable than the Karachi ones.

(B. Com., Alld., 1944).

(16) The following table gives the index numbers of industrial production of Great Britain and the number of registered unemployed persons in the same country during the years 1924-1931:—

Year	Industrial Production (Index Number)	Number of registered Unemployed (Hundred thousand)
1924	100	11.3
1925	102	12.4
1926	104	14.0
1927	107	11.1
1928	105	12.3
1929	112	12.2
1930	103	19.1
1931	94	26.4

Calculate co-efficient of correlation between production and the number of unemployed.

(B. Com., Luck., 1944).

(17) Find the co-efficient of correlation between marks obtained by candidates at an examination in two subjects A and B from the following data:—

Subject A Max. 50	Subject B—Maximum 50					Total
	11—15	16—20	21—25	26—30	31—35	
1—5					1	1
6—10	1	1	8	7	1	18
11—15	1	2	4	14	4	25
16—20			7	13	6	26
21—25			2	4	1	7
26—30			1			1
31—35				1		1
Total ..	2	3	22	39	13	79

(B. Com., Bom., 1936).

(18) The table below gives the frequency, according to age groups, of marks obtained by 65 students in a General Knowledge test. Measure the degree of relationship between age and General Knowledge.

Test marks.	Age in years.			
	19	20	21	22
200—250	4	4	2	1
250—300	3	5	4	2
300—350	2	6	8	5
350—400	1	4	6	8

(B. Com., Alld., 1947).
(M.A., Punjab, April, 1951).

(19) The following table gives the number of candidates getting different marks in Economics and Politics:—

Number of candidates getting					
Marks in Economics	Marks in Politics				
	Below 33	33—48	48—60	60 and above	Total
Below 33	3	1	1	..	5
33—48	2	6	1	2	11
48—60	1	2	2	1	6
60 & above	..	1	1	1	3
Total ..	6	10	5	4	25

Calculate the coefficient of correlation between marks obtained in the above two subjects.

(M.A., Patna, 1943).

(20) The following table gives the number of students having different heights and weights:

Height in inches	Weight in pounds			
	90—100	100—110	110—120	120—130
50—55	4	7	5	2
55—60	6	10	7	4
60—65	6	12	10	7
65—70	3	8	6	3

Do you find any relation between height and weight? How far do you consider your conclusion to be significant.

(B. Com., Alld., 1949).

(21) Calculate the coefficient of correlation and the probable error from the following table:—

$x \downarrow$ \ $y \rightarrow$	0—5	5—10	10—15	15—20	20—25	Total
0— 4	1	2				3
4— 8		4	5	8		17
8—12			3	4		7
12—16				2	1	3
Total ..	1	6	8	14	1	30

(B. Com., Alld., 1951).

(22) The following table gives the marks secured by candidates at a competitive examination held by the Federal Public Service Commission in November 1938. Examine how far success at the viva-voce test is a true index of the ability of the candidates at the written examination.

Marks secured in written examination	Marks secured at viva voce test									Total
	0-20	20-40	40-60	60-80	80-100	100-120	120-140	140-160	160-180	
800-850	..	..	..	..	..	1	..	..	..	1
750-800	..	..	..	2	1	..	..	2	..	5
700-750	..	..	2	2	3	4	..	..	1	12
650-700	..	..	1	9	3	2	3	1	1	20
600-650	..	..	5	8	11	3	2	2	1	32
550-600	..	1	10	14	4	5	1	1	..	36
500-550	..	..	1	8	4	4	..	3	1	21
450-500	..	1	8	4	3	4	1	..	..	21
400-450	1	..	3	5	1	..	1	..	1	12
350-400	..	..	3	2	1	..	..	..	..	6
300-350	..	..	1	..	..	..	..	..	..	1
Total	1	2	34	54	31	23	8	9	5	167

(M.A., Patna, 1940).

(23) The following table gives the height and weight measurements of 150 school children:

Weight in pounds	Height in inches.						Total
	40—43	44—47	48—51	52—55	56—59	60—63	
35— 54	2	20	30	..	..	..	52
55— 74	..	..	12	44	6	..	62
75— 94	..	..	..	4	16	4	24
95—114	..	..	..	..	2	4	6
115—134	..	..	..	2	..	..	2
135—154	..	..	..	..	2	2	4
Total ..	2	20	42	50	26	10	150

Calculate the coefficient of correlation between heights and weights of children.

(M.A., Alld., 1950).

(24) Find the coefficient of correlation between the speed of a motor car and the distance travelled after the brake is applied before it is put to a stop, from the following data.

Speed (miles per hr.)	Distance travelled in feet									
	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	Total
0-5	1	1	1							3
5-10		2	3	2						7
10-15		1	3	4	2					10
15-20			2	5	5	2	1			15
20-25				4	6	6	4	3		23
25-30				1	7	5	3	2		18
30-35					2	4	4	2		12
35-40						2	4	1	1	8
40-45							2	1	1	4
Total	1	4	9	16	22	19	18	9	2	100

(M.A., Patna, 1944).

(25) What are the assumptions upon which the Pearsonian co-efficient of correlation is based? How does the positive correlation differ from the negative ?

Compute the co-efficient of correlation of the short-time oscillations from the following data:—

Year		Supply	Price
1921	..	80	146
1922	..	82	140
1923	..	86	130
1924	..	91	117
1925	..	83	133
1926	..	85	127
1927	..	89	115
1928	..	96	95
1929	..	93	100

(Assume a three-year cycle, and ignore decimals).

(M. Com., Alld., 1943).

(26) Find the co-efficient of correlation from the following table.

$y=$ \ $x=$	5	10	15	20	25	30	Total
10	..	1	1	2	8	12	24
15	1	2	5	9	80	11	108
20	2	15	42	98	36	8	201
25	5	20	51	37	10	2	125
30	8	16	8	5	4	1	42
Total ..	16	54	107	151	138	34	500

(M.A., Cal., 1937).

(27) What is meant by correlation ? Give the general rules for interpreting its co-efficient.

Calculate the co-efficient of correlation between the ages of 100 husbands and wives from the following data:—

Age of Husbands in years.	Age of wives in years.					Total
	10—20	20—30	30—40	40—50	50—60	
15—25	6	3	..	..	..	9
25—35	3	16	10	..	..	29
35—45	..	10	15	7	..	32
45—55	..	..	7	10	4	21
55—65	..	..	..	4	5	9
Total ..	9	29	32	21	9	100

(M. Com., Alld., 1944 & 1950).

(28) What do you understand by co-efficient of concurrent deviations ?

Calculate the coefficient of concurrent deviations from the following data:—

n or number of pairs of observations$=47$.

c or number of pairs of concurrent deviations$=16$.

(29) Calculate the co-efficient of correlation between infant mortality and overcrowding from the following data:—

Infant mortality	109	122	96	142	151	124	125	102	109	156	122
Percentage of population over-crowded	14.9	6.3	5.8	12.2	33.2	13.3	14.6	8.8	14.9	39.8	6.3

(B. Com., Hons., Andhra 1943).

(30) From the following table calculate the co-efficient of correlation between the ages of husbands and wives and estimate the probable error of the result obtained.

Age of Husbands	Age of wives								Total
Years	15	20	25	30	35	40	45	50	
55—60						6	5	3	14
50—55					8	10	6	2	26
45—50				2	13	8	4		27
40—45				14	18	3			35
35—40			15	20	8				43
30—35	6	12	25	16					59
25—30	15	26	20	1					62
20—25	22	10	2						34
Total ..	43	48	62	53	47	27	15	5	300

(B. Com., Nagpur, 1943).

(31) Ten competitors in a beauty contest are ranked by three judges in the following orders:—

First Judge	..	1, 6, 5, 10, 3, 2, 4, 9, 7, 8.
Second Judge	..	3, 5, 8, 4, 7, 10, 2, 1, 6, 9.
Third Judge	..	6, 4, 9, 8, 1, 2, 3, 10, 5, 7.

Use the rank correlation co-efficient to discuss which pair of judges have the nearest approach to common tastes in beauty.

(M.A., Alld., 1952).

(32) Calculate the coefficient of correlation of the short time oscillations for the following indices of supply and price of a community:—

Year	Index of Supply	Index of Price	Year	Index of Supply	Index of Price
1	186	222	9	154	204
2	168	228	10	192	192
3	156	232	11	210	186
4	166	216	12	216	184
5	178	200	13	204	190
6	186	196	14	194	196
7	154	208	15	234	182
8	136	214			

(M.A., Raj., 1951).

(33) Why is it of little avail to compute the coefficient of correlation from pairs of actual items in two sets of data exhibiting both cyclic and trend variability ?

From the standard deviation cycles of the production and price of coal for the period 1900—1910 find the coefficient of correlation:—

Year		Production σ	Price σ
1900	..	−0.06	+0.12
1901	..	−0.85	−1.56
1902	..	+0.79	−0.32
1903	..	+1.06	+0.16
1904	..	+0.94	+2.79
1905	..	+2.40	−0.20
1906	..	+0.24	+0.12
1907	..	+0.37	−0.16
1908	..	−1.03	−0.72
1909	..	+0.43	−0.16
1910	..	+0.46	−0.12

(M. Com., Alld., 1946).

(34) Draw Correlation Graph of the following two series, and also compute their Coefficient of Correlation. Can you tell by inspection of the data what the answer should be ?

x	y
60	−180
70	−210
−30	90
−50	150
60	−180
−50	150
−40	120
−20	60

(35) Compute Coefficient of Correlation by Karl Pearson's method and also by the method of rank differences in respect of the pairs of series in (*a*) and (*b*) given below. Do the two methods give like results with these data ? Why ?

(*a*) x	y	(*b*) x	y
100	400	11	48
200	600	7	46
300	700	12	54
400	500	13	56
500	100	10	52
600	300	9	44
700	200	8	50

(36) From the data given in (*a*) and (*b*) of exercise 35 above, draw two scatter diagrams separately and state the indications the diagrams give regarding the nature and degree of correlation in each case.

CHAPTER XVI

REGRESSION.

If we are able to establish the fact of correlation between two variables, it will be our natural curiosity to know the extent to which one variable varies in response to a given variation in the other variable. In other words, we will be curious to know that if the subject changed by 1%, by what fraction of 1% would the relative change. This means we will be interested to know the ratio of variation.

Ratio of Variation.

Although almost perfect correlation may exist between two series x and y, yet the proportional movements may be very different. Price and supply of goods may, for instance, have a high positive correlation, almost approaching unity, yet the movement of price may be proportionally greater than that of supply. In many cases a measure of this proportional variation can be employed very usefully. This measure is furnished by the ratio of variation.

Ratio of variation is the average ratio of the percentage deviations from the mean in the relative (y) as compared with those in the subject (x). That series in which the average percentage deviations are larger is generally taken as the subject, so that the ratio of variation may be less than unity. The ratio of variation can be found out in two years: (1) mathematically and (2) graphically.

Mathematical computation of Ratio of Variation.

According to this method, the ratio of variation can be calculated as follows:—

(1) Compute the deviation of the relative from the mean for each date; that is, compute the percentage movement of the relative from the mean.

(2) Similarly, compute the percentage movement of the subject from the mean.

(3) Divide the percentage deviation of the relative by the corresponding deviation of the subject.

(4) Add up the quotients and divide the summation by the number of quotients; the resulting figure is the ratio of variation.

Thus, the procedure involves the measurement of the proportional variation for both the series and then comparison by means of a ratio. This mathematical method is, however, not considered suitable for most social and economic series inasmuch as such series are generally irregular. The graphical method is instead preferred.

Galton Graph.

Ratio of variation can be measured graphically by means of the Galton Graph. This graph was devised by Francis Galton and is named after him. It is particularly useful when the oscillating deviations are irregular. The graph may be drawn thus for any two related variables, x and y:—

(1) Convert the series of original data into series of index numbers, using the average of the values of the items of each series as the base.
(2) Allow for time-lag, if any.
(3) Plot the indices of the subject on the vertical scale.
(4) Plot the indices of the relative on the horizontal scale.
(5) The resulting diagram will look like a "gunshot" chart, or a scatter diagram as in Figure 30A. Draw the line of best fit through the dots on the diagram, as was done in Figure 30A. This line should follow the general trend of the plotted points. In drawing this line the following conditions should be observed:
   (a) There should be approximately an equal number of points on either side of the line;
   (b) The points on either side of the line should be equi-distant from the line; and
   (c) The line should pass through the averages of x (subject) and y (relative) series. In the case of the index number series framed according to (1) above, average of $x=100$ and that of $y=100$. The line, in the case of Galton Graph, should, therefore, pass through 100, 100.

(6) The slope of the line, that is, the tangent of the angle between y and x will give the ratio of variation between y and x. The reverse of this result will give the corresponding ratio of variation between x and y.

Reading the Galton Graph.

The following rules will help in reading a Galton Graph:—

(1) If correlation is perfect, all the dots on the graph will lie on a straight line. For example:

x: 1, 2, 3, 4, 5, 6, 7, 8

y: 2, 4, 6, 8, 10, 12, 14, 16

(2) If the line slopes upwards from left to right, the correlation is direct; that is, the variables move in sympathy.

(3) If the line slopes to the right, the correlation is inverse; that is, the variables move in opposition.

(4) If both the subject and the relative vary by the same percentage, say, both increase (or decrease) by 2%, the ratio of variation is obviously unity, and the line drawn through the plotted points will be a line at 45° to the horizontal. Such line at 45° is called the Line of Equal Proportional Variation, or simply the Line of Equal Variation. See Figure 31.

(5) If the subject changes more proportionally than does the relative, the line of best fit will be at an angle of more than 45° to the horizontal. If, on the other hand, the relative changes more proportionally than the subject, the line will be at less than 45° to the horizontal. The lines at more than or less than 45° to the horizontal are termed Lines of Regression.

(6) The greater the divergence of the line of regression from the line of equal variation, the less is the correlation, and *vice versa*. See Figure 31.

(7) To find out the ratio of variation, a horizontal line should be drawn at any point cutting through the line of regression. See Figure 31. The distance

from the point of intersection (C) to the ordinate axis (B), divided by the distance between B and the point at which line of regression touches the ordinate axis (A) will give the required ratio of variation. In other words, ratio of variation= tangent of angle BAC.

Example 1. Required to find out the ratio of variation by means of the Galton Graph, between x and y given in col. (2) and (4) of Table 63. (The data are hypothetical).

Table 63. *Original Series x and y and their index numbers, for measuring Ratio of Variation.*

Year	x	Index of x (20=100)	y	Index of y (200=100)
(1)	(2)	(3)	(4)	(5)
1	12	60	140	70
2	20	100	180	90
3	18	90	180	90
4	25	125	190	95
5	18	90	210	105
6	13	65	190	95
7	12	60	160	80
8	17	85	140	70
9	26	130	200	100
10	26	130	260	130
11	32	166	330	165
12	28	140	220	110
13	24	120	260	130
14	20	100	220	100
15	16	80	180	90
16	13	65	140	70
$n=16$	$\Sigma m=320$		$\Sigma m=3{,}200$	

$$\text{Average of } x \text{ series} = \frac{\Sigma m}{n} = \frac{320}{16} = 20$$

$$\text{Average of } y \text{ series} = \frac{\Sigma m}{n} = \frac{3200}{16} = 200$$

With the average of each series as the base, the original series have been converted into index number series in columns (3) and (5), respectively. The standard deviation of the index series x works out to be greater than that of index series y, while their means are equal, 100 each. Therefore, the variability in x is greater than that in y. The index series x is consequently taken as the subject and the index series y as the relative. The index of x is ranged along the vertical scale in Figure 31 and the index of y along the horizontal scale. The line of regression AD is then drawn, free hand, keeping in view the conditions laid down in para. (5) while explaining the Galton Graph. The line passes through x=100, y=100. A point B is then chosen at the ordinate axis, and a line parallel to the horizontal is drawn from that point to cut the line of regression at C. BC read on horizontal scale is 45 and BA read on the vertical scale is 69.

$$\text{Ratio of Variation} = \frac{BC}{BA} = \frac{45}{69} = 0.65$$

Now we can say that for every change of 1% in the subject x, the relative y would tend to change by 0.65%, or that the proportional changes in the subject are greater than the proportional changes in the relative. Thus, in our example x fluctuates more widely than y.

We have taken hypothetical data in Table 63. In actual practice, x may correspond to price and y to supply of goods, or x to industrial shares and y to industrial production, or x to rainfall and y to yield of crop.

Ratio of Regression.

The difference between the ratio of variation and unity gives the Ratio of Regression. In our example, Ratio of Regression is equal to (1−0.73) or 0.27.

Usefulness of Galton Graph.

The Galton Graph helps us to find out whether there is any correlation between two given variables, and, if yes, whether the correlation is positive or negative, perfect or limited. The

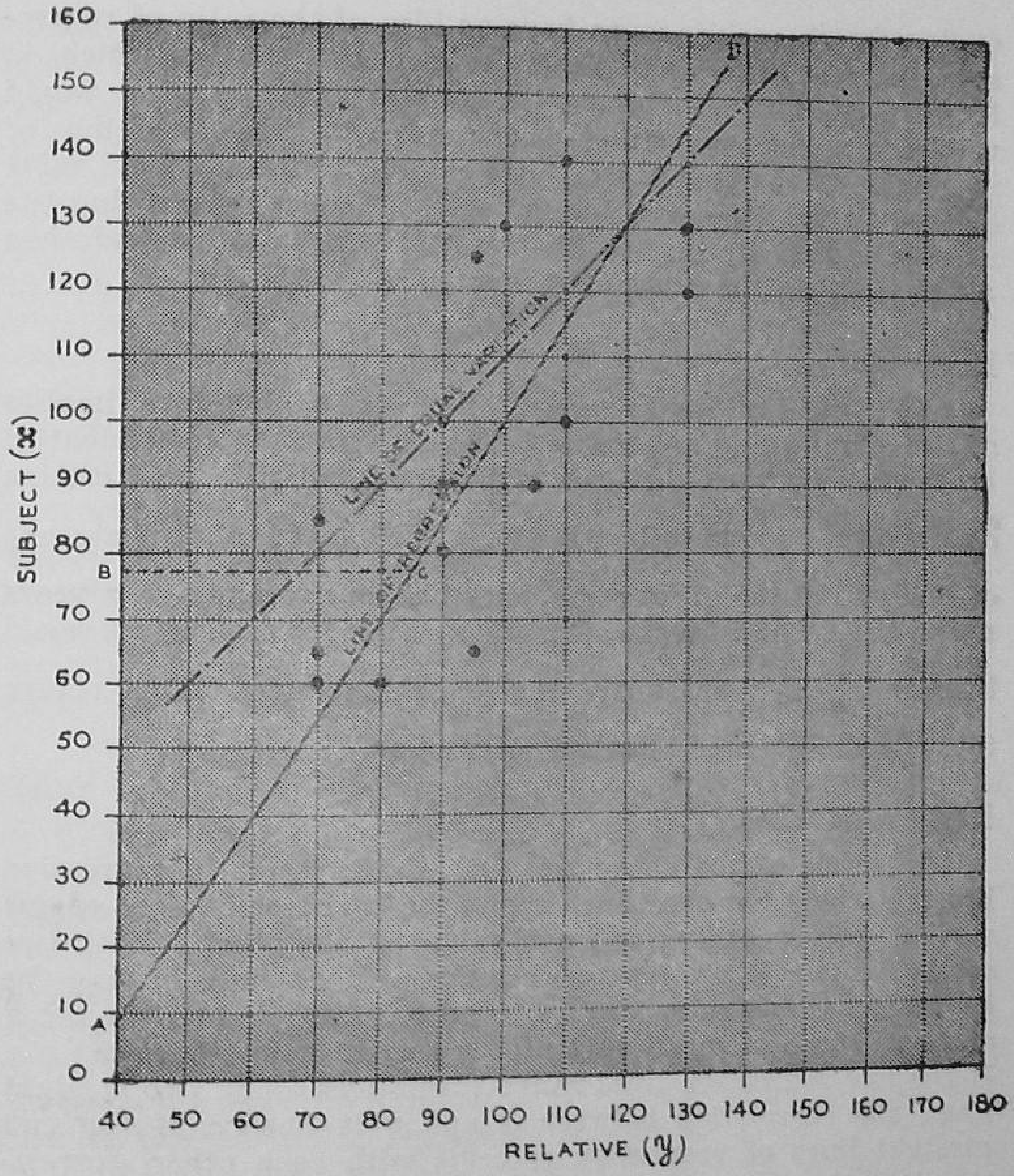

SUBJECT (x)
RELATIVE (y)
LINE OF EQUAL VARIATION
LINE OF REGRESSION
A
B
C
D
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
40 50 60 70 80 90 100 110 120 130 140 150 160 170 180

Fig. 31.

Galton Graph enables us to have an idea of the ratio of regression. It also helps us to know the ratio of variation which, in its turn, tells us the proportion by which the relative would vary for a given variation in the subject. From the line of regression in the Galton Graph we can locate the expected value of the relative for a given value of the subject by drawing the necessary ordinate on the line from the given value and then reading out on the relevant scale.

Regression.

To regress means to return. Regression, therefore, implies an act of returning or passing back. Suppose the co-efficient of correlation between the ages of husband and wife in a certian community is found to be 0.8, or $\frac{4}{5}$. This means that if a group of husband in the given community has an average age x years above the general average, the average age for their wives would be only $\frac{4}{5}x$ years above the general average. This return towards the average is called Regression.

Regression Line.

In terms of the above meaning of regression, a regression line is nothing but one which shows the best mean values of one variable corresponding to mean values of the other. With two series x and y, there are two regression lines, one showing the best mean values of x corresponding to mean y's and the other showing the best mean values of y corresponding to mean x's. If the two regression lines range themselves along two straight lines, the correlation between x and y is linear. If the two straight lines of regression coincide with each other, correlation is perfect. If they do not coincide, the correlation is limited. The farther are the lines scattered from each other, the less is the degree of correlation. If they cut each other at right angles, correlation is zero.

Regression line is, in effect, a straight line trend fitted to the data, as plotted, by the method of least squares. The essential difference between the line of best fit by the method of least squares and the line of regression is that, whereas the

former describes the average change in a given series accompanying a unit change in *time*, the latter similarly describes the average relationship between *two variables*. See Figure 28A in the Chapter on Analysis of Time Series. The line of best fit in that Figure passes through the arithmetic means of x and y, of which one is time. The line of best fit drawn freehand in the Scatter Diagram in Figure 30A, and the line of regression, also drawn freehand, in Figure 31 both give the average relationship between two variables. They pass through the means of their respective series, and have equal number of dots on either side. Thus, a regression line is a line which is so drawn that the distance of the dots from that line in the Scatter Diagram is minimum on the average.

The functions of regression line are several. It tells us whether correlation does or does not exist between x and y. It also tells us whether the correlation, if it exists, is positive or negative, high or low. Its main utility, however, is that it enables the expression of the probable movement of one variable in terms of the other.

How to draw the regression lines? How to get the best mean values of x and y ? This is done with the help of regression equations.

Regression Equations.

A regression equation is the algebraical expression of regression line. There are two regression equations: one to measure the best mean values of x corresponding to mean y's, that is, to measure the regression of x on y; and, the other to measure the best mean values of y corresponding to mean x's, that is, to measure the regression of y on x.

Regression of x or y

$$x-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y})\ldots\ldots\ldots\ldots(1)$$

Regression of y on x

$$y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x})\ldots\ldots\ldots\ldots(2)$$

where, $\bar{x}$ and $\bar{y}$ are the means of x and y series respec-

pectively, σ_x and σ_y their respective standard deviation and r the co-efficient of correlation between x and y.

Regression equations serve two purposes. Firstly, they furnish the values for the regression lines and secondly, they provide a numerical method of finding out the most probable value of x for a given y and the most probable value of y for a given x.

Example 2. The ages of husband and wife in a community were found to have a correlation co-efficient equal to $+0.8$; the average of husbands' ages was 25 years and that of wives' ages 22 years; their standard deviations were respectively 4 and 5 years. Required to draw the two lines of regression and from the lines to measure (i) the expected age of husband when wife's age is 12 years and (ii) the expected age of wife when husband's age is 33 years.

The given values are as under:—

$$\bar{x}=25 \text{ yrs.}, \qquad \sigma_x=4 \text{ yrs.}$$
$$\bar{y}=22 \text{ yrs.}, \qquad \sigma_y=5 \text{ yrs.}$$
$$r=+0.8$$

We are required to find (i) the expected value of x when y is 12 years and (ii) the expected value of y when x is 33 years.

(i) Regression of x on y:—

$$x-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

$$\text{or, } x-25=0.8\times\frac{4}{5}(y-22)$$

$$\text{or, } x-25=0.64y-13.68$$

$$\therefore x=0.64y+11.32$$

Thus the equation of the line showing regression of x on y works out to be

$$x=0.64y+11.32$$

Now,

if $y=1$, $x=0.64+11.32=11.96$

if $y=10$, $x=6.40+11.32=17.72$, and so on.

Thus,

x :	17.72	20.92	24.12	27.32	30.52
y :	10	15	20	25	30

The above ordinates are plotted in Figure 32. We get the regression line AB showing regression of x on y. This line will enable us to interpolate the values of x for given values of y. To find the expected value of x when y is 12, a dotted line parallel to OY is drawn from the point 12 on the horizontal scale. This dotted line cuts the line of regression at C. From C a line parallel to the horizontal base is drawn to cut OY at D. DO read from the vertical scale gives $x=19$. We may thus conclude that if in the given community the age of a wife is 12 years, the age of her husband is most likely to be 19 years.

(ii) Regression of y on x:—

$$y-\bar{y}=r\frac{\sigma_x}{\sigma_y}(x-\bar{x})$$

$$\text{or, } y-22=0.8\times\frac{5}{4}(x-25)$$

$$\text{or, } y-22=x-25$$

$$\therefore\ y=x-3$$

Thus the equation of the line showing regression of y on x works out to be

$$y=x-3$$

Now,

x :	10	15	20	25	30
y :	7	12	17	22	27

These ordinates give the line MN which shows the regression of y on x. From the line, it is found out by drawing the required ordinate PQ that the expected age of wife corresponding to a husband of 33 years is 30 years.

Best values of x for given values of y and best values of y for given values of x can also be found out straight from the regression equations instead of through the regression lines.

Example 3. Required to estimate (i) the age of husband when wife's age is (a) 12 years and (b) 22 years, and (ii) the age of wife when husband's age is (c) 28 years and (d) 33 years.

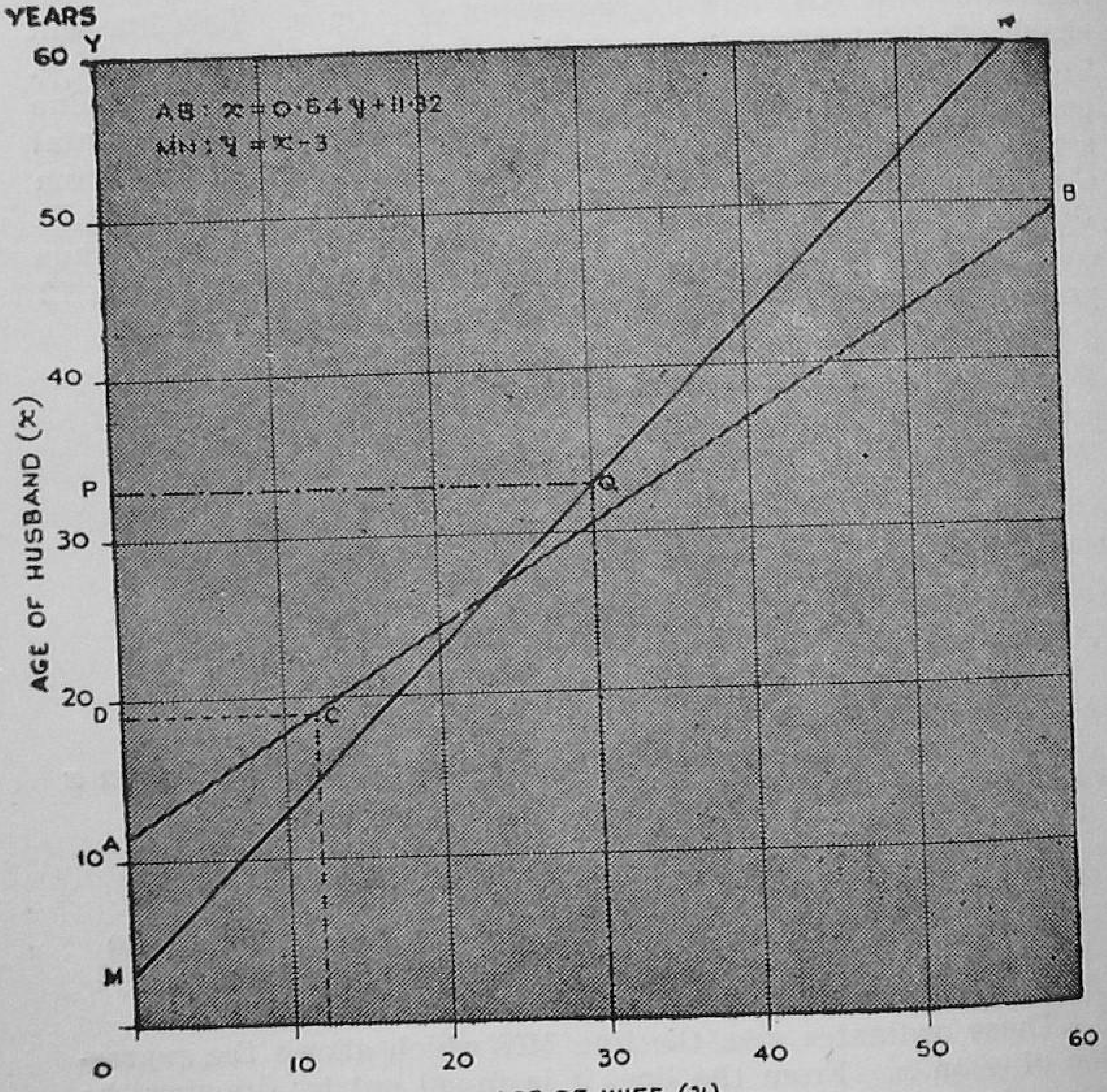

YEARS
Y
60
50
40
30
20
10
P
D
A
M
Q
C
B
N
AB : x = 0·64 y + 11·82
MN : y = x − 3
AGE OF HUSBAND (x)
0
10
20
30
40
50
60
YEARS
AGE OF WIFE (y)

Fig. 32.

(i) Regression of x on y:

$$x-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

$$\text{or, } x-25=0.8\times\frac{4}{5}(y-22)$$

$$\therefore\ x=0.64\ y+11.32$$

Now,

(*a*) when $y=12$, $x=\ 7.68+11.32=19.0$ years,

(*b*) when $y=22$, $x=13.08+11.32=24.4$ years.

(ii) Regression of y on x:

$$y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

$$\text{or, } y-22=0.8\times\frac{5}{4}(x-25)$$

$$\therefore\ y=x-3$$

Now,

(*c*) when $x=28$, $y=28-3=25$ years,

(*d*) when $x=33$, $y=33-3=30$ years.

It will be seen that the values arrived at in (*a*) and (*d*) above are the same as those arrived at with the help of regression lines in example 2. This correspondence is natural, since the regression equations used in both the methods are the same.

Co-efficient of Regression.

The co-efficient of regression gives the value by which one variable increases for a unit increase in the other variable. Like the equations of regression, there are two equations for co-efficients of regression.

Regression co-efficient of x on $y=r\dfrac{\sigma_x}{\sigma}$ ....(1)

Regression co-efficient of y on $x=r\dfrac{\sigma_y}{\sigma}$ ....(2)

The regression co-efficient of x on y gives the value by which x variable increases for a unit increase in y. Likewise, the regression co-efficient of y on x gives the value by which y variable increases for a unit increase in x.

Example 4. Required to compute the coefficient of regression of x on y and of y on x from the data given in example 2.

$$\text{Coefficient of regression of } x \text{ on } y = r\frac{\sigma_x}{\sigma_y}$$

$$= 0.8 \times \frac{4}{5} = 0.64$$

$$\text{Coefficient of regression of } y \text{ on } x = r\frac{\sigma_y}{\sigma_x}$$

$$= 0.8 \times \frac{5}{4} = 1$$

We may now say that x variable increases by 0.64 for a unit increase in y and that y variable increases by unity for a unit increase in x.

It will be seen that the above values of the coefficients provide the slope of the regression lines in example 2. Thus, the coefficient of regression gives the slope of the regression line. And, it is the slope of the straight line which measures the average increase in one variable corresponding to a unit increase in the other.

Coefficient of correlation between two series x and y is the square-root of the product of their coefficients of regression.

$$\text{Since, } r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x} = r^2$$

$$\therefore\ r = \sqrt{r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x}}$$

Example 5. Required to estimate the value of coefficient of correlation between x and y if the two coefficients of regression are 0.64 and 1 (*vide* example 4).

$$r = \sqrt{0.64 \times 1}$$

$$= 0.8$$

General Remarks.

It must have been noticed that whereas ratio of variation is invariably measured from index number series, coefficient of regression can be measured from the original series. Also, while graphic method alone is recommended for studying the ratio of variation, regression can be studied both graphically and numerically—graphically by means of regression lines and numerically through the coefficient of regression and regression equations.

The usefulness of the study of regression should by now be clear. How a line of regression helps in determining the existence of correlation between two variables, as also in determining whether it is high or low, positive or negative, was emphasized in the Chapter on Correlation while dealing with the Scatter Diagram. In this Chapter, further uses of the regression line have been pointed out: it can be used to measure the ratio of variation as also to measure the best value of one variable for a given value of the other variable. This is an invaluable service which the line of regression performs. It is good to know the regression of one variable on the other. In actual practice, one very often comes across the need of estimating the probable value of one variable for a given value of the other, when the two variables are known to be correlated. If the rainfall were of a particular order, what the likely yield of the crops would be; if the wages were at a particular level, how they would react on production; these and similar other estimates are very often required and have to be made. Knowledge of regression analysis comes in handy in framing these estimates. These estimates are, however, significant only when the coefficient of correlation has a high value and when the correlation, as measured, is assumed to have a general applicability. Such conditions are not commonly met with in social and economic statistics, and, therefore, we come back to the idea that correlations should generally be regarded as measurements merely of specific inter-relationships, rather than of general laws.

The terms line of regression, regressions equation and coefficient of regression date back to early studies made by Francis Galton regarding the relation between the heights of fathers and heights of sons. Galton found that on an average

sons deviated less from the mean heights of the race than their fathers. Whether the fathers were below or above the average, the sons tended to go back or regress towards the average. He, therefore, named the line which graphically described the average relationship between these two variables as the line of regression. The term is now used generally, and so are used its other two companions—the equation and the coefficient of regression.

EXERCISES

(1) Explain the concepts of Regression and Ratio of Variation and state their utility in the field of economic enquiries.

(M.A., Pb., Sept., 1952).

(2) Plot a Galton graph from the following table, and show the Ratio of Variation between Bank Clearings and Immigrants for eight years.

Year	Immigrants (tens of thousands)	Bank clearings (millions of £)
1	79	49
2	52	40
3	33	25
4	55	35
5	46	35
6	62	34
7	31	34
8	34	28
Average	49	35

(M. Com., Alld., 1947).

(3) From the following data ascertain the ratio of variation between sales and profits:—

Year		Sales (Lakhs of Rs.)	Profits (Lakhs of Rs.)
1940	..	72	42
1941	..	84	52
1942	..	66	48
1943	..	60	46
1944	..	48	30
1945	..	42	28
1946	..	54	36
1947	..	63	38
1948	..	51	34
1949	..	57	42
1950	..	69	44
1951	..	54	40
Total	..	720	480
Average	..	60	40

(M.A., Pb., April, 1952).

(4) The following table gives hypothetical data about price and supply of a commodity for January to November:—

		Price (x)	Supply (y)
January	..	15	60
February	..	22	30
March	..	40	44
April	..	26	80
May	..	34	52
June	..	58	68
July	..	50	116
August	..	45	100
September	..	70	90
October	..	60	140
November	..	55	120

Draw a Galton Graph and determine the ratio of variation. What light does the resulting ratio throw on the nature and extent of correlation between x and y?

[y series shows a time-lag of one month. This lag should be allowed for after converting both the series into index numbers.]

(5) Calculate Karl Pearson's coefficient of correlation and the regression equations from the following data:—

Age of husband	Age of wife
18	17
19	17
20	18
21	18
22	18
23	19
24	19
25	20
26	21
27	22

(M. Com., Alld., 1951).

(6) (*a*) Given:—

	X series	Y series
Mean	18	100
Standard deviation ..	14	20

Coefficient of correlation between x and y series is $+0.8$.

Find out the most probable value of y if x is 70 and most probable value of x if y is 90.

(*b*) if two regression coefficients are 0.8 and 0.6, what would be the value of the coefficient of correlation.

(M. Com., Alld., 1952).

(7) The following data are given for marks in subjects A and B in a certain examination:—

Mean marks in A	=	39.5
Mean marks in B	=	47.5
Standard Deviation marks in A ..	=	10.8
Standard Deviation marks in B ..	=	16.8
Coefficient of correlation between A and B	= +	0.42

Draw the two lines of regression, and explain why there are two equations of regression.

Also give the expectations of marks in subject B for candidates who secured 50 marks in subject A.

(M.A., Pb., Sept., 1951).
(M. Com., Alld., 1946).

(8) A study of wheat prices at Hapur and Kanpur yields the following data:—

		Hapur	Kanpur
		Rs.	Rs.
Average price per maund	..	2.463	2.797
Standard deviation	..	0.326	0.207

Coefficient of correlation between two prices +0.774

Estimate from the above data the most likely prices of wheat (*a*) at Hapur corresponding to the price of Rs. 2.334 at Kanpur, (*b*) at Kanpur corresponding to the price of Rs. 3.052 at Hapur.

(M. Com., Alld., 1950).

(9) (*a*) What is meant by coefficient of concurrent deviations? Calculate the coefficient from the following data:—

Number of pairs of concurrent deviations = 198;
Number of pairs of observations = 340.

(*b*) Given the following data, calculate the expected value of y when x is 12.

		x	y
Average	..	7.6	14.8
Standard deviation	..	3.6	2.5

$r = +\ 0.99$

(M. Com., Alld., 1945).
(M.A., Raj., 1951).

(10) Two lines of regression are given by $x+2y-5=0$ and $2x+3y-8=0$ and $\sigma_x{}^2=12$. Calculate the value of $\bar{x}$, $\bar{y}$, $\sigma_x{}^2$, and r.

(M.A., Alld., 1952).

(11) Find the coefficient of correlation between the ages of husbands and wives at marriage from the following table.

Ages of husbands and wives at marriage

Husband's age (years)	Wife's age (years)
23	18
27	20
28	22
28	27
29	21
30	29
31	27
33	29
35	28
36	29

Also find its equation of regression in its simplest form.
(M.A., Pb., 1950).

(12) Given: $\bar{x} = 5.5$ $\sigma_x = 1.58$

$\bar{y} = 5.5$ $\sigma_y = 1.64$

$r = +0.95$

Draw the two regression lines and from them measure the value of y when x is 18 and the value of x when y is 25. Verify your result by calculating the required values of x and y from the regression equations straight.

(13) From the x and y series given below calculate the values required for computing the two coefficients of regression, and calculate the coefficients.

From the values of the coefficients of regression, estimate the coefficient of correlation and check up if it tallies with the coefficient of correlation computed directly.

x	y
78	125
89	137
97	156
69	112
59	107
79	136
68	123
57	108

(14) Given:

$\bar{x}=30$ $\Sigma d^2_x = 164$

$\bar{y}=25$ $\Sigma d^2_y = 138$

$r=+0.81$ $n = 10$

Find the value of y when x is 100 and the value of x when y is 16, and verify your result graphically.

(15) From the following data draw the two regression lines and from them read the nature and extent of correlation between x and y:—

Year	x	y
1	12	38
2	16	20
3	18	17
4	19	21
5	19	38
6	18	38
7	17	40
8	11	42
9	9	46
10	11	40

CHAPTER XVII

ASSOCIATION OF ATTRIBUTES.

Statistics of Attributes.

In Chapter VII[1] reference was made to 'descriptive' and 'numerical' characteristics of statistical data, and characteristics like sex, civil condition, caste, religion and infirmity were mentioned as descriptive characteristics. Quantitative character of statistical data arises in two ways. First, the investigator may note only the *presence* or *absence* of some attribute in a series of objects individuals, and *count* the number of those who possess it and of those who do not. For instance, in a given population the number of the deaf and not-deaf or of the sane and insane may be counted. In such cases, the quantitative character arises solely in the process of counting. Second, the investigator may note or measure the actual *magnitude* of some variable character for every one of the objects or individuals observed. For instance, height of students in a class, length of leaves of a certain tree, or prices of certain commodities may be *recorded.* Such records are quantitative in character. In these cases, therefore, the observations themselves are quantitative in character.

As pointed out in Chapter VII, the first kind of observations are based on descriptive characteristics and are termed **Statistics of Attributes,** while the second are based on numerical characteristics and are called **Statistics of Variables.** So far, in all the chapters, we have been concerned with statistics of variables. We have studied how variables are analyzed, compared, and correlated with one another. It is now proposed to study how relationship can be established between two attributes, and how that relationship can be discovered by the method of Association.

Notation and Terminology.

While discussing classification of data according to attributes in Chapter VII it was pointed out that when one attribute is noticed, two distinct classes are formed. These two classes,

[1] *Cf.* page 65.

however arbitrary their boundary, are exclusive of each other. Such classification was, in that Chapter, denominated as simple classification or classification by dichtomy.

To discuss the theory of association and its application in practice it is necessary to have some simple notation for the classes formed and for the measurements assigned to each of them. Accordingly, we shall use the capital letters A, B, C,.... to denote the several **attributes.** An object or individual *possessing* the attribute A will be termed simply A, that possessing B, B. The class whose members *possess* the attribute A will be termed the **class** A. Similarly, we shall use the small letters *a, b, c,*...... (generally, the Greek letters α, β, γ, ... are used) to denote the *absence* of the attributes A, B, C,........ Thus, if A represents the attribute blindness, *a* represents sight, *i.e.,* non-blindness; if B stands for insanity, *b* stands for sanity. **Combinations** of attributes will be represented by grouping together the letters that indicate the attributes concerned. Thus, if A represents blindness and B insanity, AB represents blindness and insanity. If the presence and absence of these attributes are noticed, then

Combination	AB	stands	for	blindness and insanity
,,	A*b*	,,	,,	blindness and sanity
,,	*a*B	,,	,,	sight and insanity
,,	*ab*	,,	,,	sight and sanity,

and, similarly, the class AB includes all those who are blind and insane, the class A*b* all those who are blind and sane, and so on. If a third attribute be noted, for example, deafness, and denoted by C, the class ABC includes those who are at once blind, insane and deaf, and the class AB*c* those who are blind and insane but not deaf.

The number of observations assigned to any class will be termed the frequency of the class, or briefly **class-frequency.** Class-frequencies will be denoted by placing the corresponding class-symbols in brackets. Thus,

(A) denotes number of A's, *i.e.,* objects possessing attribute A.

(A*b*) denotes number of A*b's, i.e.,* objects possessing attribute A but not B, and so on.

The attributes denoted by capitals A B C......may be termed **positive attributes,** and their contraries denoted by small letters **negative attributes.** Thus the class A, AB, ABC are positive classes ; the classes a, ab, abc, negative classes. AB and ab, Ab and aB, AbC and aBc, are pairs of **contrary classes.**

A class specifying one attribute is known as the class of first order ; while that specifying two, that of the second order. Thus, A is a class of the **first order,** AB or BC that of the **second order.** Similarly, (A), (Ab), (aBC) are class-frequencies of the first, second and third orders respectively.

The series of classes given by any one positive class and the classes whose symbols are derived therefrom by substituting small letters for one or more of the capital letters in all possible ways will be termed as **aggregate.** Thus (AB), (Ab), (aB), (ab) together form an aggregate of frequencies of the second order.

When no attributes are specified, the total number of observations constitutes the **Universe** with its limits specified, and will be denoted by the letter N.

It should now be clear that the Universe must be equal to the number of A's plus the number of a's. Similarly, the number of A's should equal the number of A's that are B plus the number of A's that are not B ; and so on. It means that any class-frequency can be analyzed into higher class-frequencies. Thus,

$$N=(A)+(a),$$

$$N=(B)+(b),$$

$$N=(AB)+(Ab)+(aB)+(ab),$$

$$(A)=(AB)+(Ab),$$

$$(B)=(AB)+(aB),$$

$$(a)=(ab)+(aB)=N-(A)$$

The classes specified by attributes of the highest order are termed the **ultimate classes** and their frequencies, the **ultimate class-frequencies.** If we know (AB) and (Ab) we can find (A); and, if an addition we know (ab) and (aB) we can not find only find (a) but also N. This is due to the fact, noted above, that every class-frequency can be expressed as the sum

of certain of the ultimate class-frequencies. Therefore, to specify the data completely, it is only necessary to know the ultimate class-frequencies. An example will further clear the point.

Example 1. Given the following ultimate class-frequencies, find the frequencies of the positive and negative classes and the whole number of observations, N :—

$(AB)=100$; $(Ab)=50$.
$(aB)=80$; (ab) 40.

To find N we must remember that it will be equal to the aggregate of the frequencies of the second order. Thus,

$$N=(AB)+(Ab)+(aB)+(ab)$$
$$=100+50+80+40$$
$$=270.$$

The frequency of any first-order class, *e.g.*, (A), is given by the total of the two second-order frequencies the class-symbols for which contain the same letter. Thus,

$$(A)=(AB)+(Ab)=100+50=150$$

$$(B)=(AB)+(aB)=100+80=180$$

$$(a)=(aB)+(ab)=80+40=120$$
or $$(a)=N-(A)=270-150=120$$

$$(b)=(Ab)+(ab)=50+40=90$$
or $$(b)=N-(B)=270-180=90$$

	A	a	N
B	(AB) 100	(aB) 80	(B) 180
b	(Ab) 50	(ab) 40	(b) 90
N	(A) 150	(a) 120	N 270

The nine-squares table given above affords an easy and quick manner of getting the required class-frequencies. If given values are filled in it, the required ones may be computed from it, as shown in the table.

If the values of any two classes in each of the equations used in the solution of the above example are known, the value of the third can be easily found. For example,

if, $(a)=(aB)+(ab)$
then $(ab)=(a)-(aB)$.

And, the expression of any class-frequency in terms of the positive frequencies is most easily obtained by a process of step-by-step substitution ; thus

$$\begin{aligned}(ab) &= (a)-(a\mathrm{B})\\ &= [\mathrm{N}-(\mathrm{A})]-[(\mathrm{B})-(\mathrm{AB})]\\ &= \mathrm{N}-(\mathrm{A})-(\mathrm{B})+(\mathrm{AB}).\end{aligned}$$

The expression of other class-frequencies in terms of positive frequencies can be made by a similar process of substitution.

Similarly, if eight ultimate frequencies of the third order are given, or sixteen ultimate frequencies of the fourth order are given, all the positive and negative class-frequencies and N can be obtained from them by mere addition.

Example 2. Given:

$$\begin{aligned}(\mathrm{ABC}) &= 149 & \qquad (a\mathrm{BC}) &= 204\\ (\mathrm{AB}c) &= 738 & (a\mathrm{B}c) &= 1762\\ (\mathrm{A}b\mathrm{C}) &= 225 & (ab\mathrm{C}) &= 171\\ (\mathrm{A}bc) &= 1196 & (abc) &= 21842\end{aligned}$$

Find the values of A, B, C, AB, AC and BC.

$$\begin{aligned}(\mathrm{A}) &= (\mathrm{ABC})+(\mathrm{A}b\mathrm{C})+(\mathrm{AB}c)+(\mathrm{A}bc)\\ &= 149+225+738+1196\\ &= 2308\\ (\mathrm{B}) &= (\mathrm{ABC})+(\mathrm{AB}c)+(a\mathrm{BC})+(a\mathrm{B}c)\\ &= 149+738+204+1762\\ &= 2853\\ (\mathrm{C}) &= (\mathrm{ABC})=(\mathrm{A}b\mathrm{C})+(a\mathrm{BC})+(ab\mathrm{C})\\ &= 149+225+204+171\\ &= 749\\ (\mathrm{AB}) &= (\mathrm{ABC})+(\mathrm{AB}c)\\ &= 149+738\\ &= 887\\ (\mathrm{AC}) &= (\mathrm{ABC})+(\mathrm{A}b\mathrm{C})\\ &= 149+225\\ &= 374\\ (\mathrm{BC}) &= (\mathrm{ABC})+(a\mathrm{BC})\\ &= 149+204\\ &= 353\end{aligned}$$

CONSISTENCE

Any class-frequencies observed within one and the same universe may be said to be consistent with one another. They do not in any way conflict with one another.

A necessary condition for the consistence of a set of independent class-frequencies is that no ultimate class-frequency should be negative. Hence, to test the consistence of given class-frequencies we should only calculate the ultimate class-frequencies, and find out whether any one is negative. If any one if found to be negative, the conclusion is that the data are inconsistent. If, on the other hand, no ultimate class-frequency is found to be negative, the data are consistent. Let us take an example.

Example 3. An investigator returned the following data: Of the 1000 students he interviewed, 811 liked tea, 752 liked coffee and 418 liked cocoa ; 570 liked tea and coffee, 456 liked tea and cocoa and 348 liked coffee and cocoa ; 297 liked all three. Show that this information, as returned, must be incorrect.

Let A stand for likeness for tea, B for coffee and C for cocoa. Then,

$$(N)=1000 \qquad (AB)=570$$
$$(A)=\ 811 \qquad (AC)=356$$
$$(B)=\ 752 \qquad (BC)=348$$
$$(C)=\ 418 \qquad (ABC)=297$$

Now,

$$(abc)=N-(A)-(B)-(C)+(AB)+(AC)+(BC)-(ABC)$$
$$=1000-811-752-418+570+356+348-297$$
$$=-4$$

The test of consistence is that no ultimate class-frequency can be negative. But, we have found that *abc* works out to be negative. The data are not consistent. If the data returned, as given in the example, are alleged to be the result of an actual enquiry in a definite universe, there must have been some misprint or mis-count or mis-reporting.

PROBABILITY AND EXPECTATION

When a coin is tossed once, it must fall heads or tails. The probability (pure chance) that it would fall heads is $\frac{1}{2}$. When a coin is tossed 50 times, the expectation of a head coming up is $\frac{1}{2}\times50=25$. Therefore **expectation** is equal to the product of probability and the number of observations. If two coins are tossed, the chance of two heads or two tails coming up is reduced to $\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$.

If two attributes A and B are studied in a Universe N and the class-frequencies of the attributes are (A) and (B), then

$$\text{Probability of (A)}=\frac{(A)}{N}$$

$$\text{Probability of (B)}=\frac{(B)}{N}$$

Probability of (A) and (B) combined $=\frac{(A)}{N}\times\frac{(B)}{N}$, and

$$\text{Expectation of (A) and (B) combined}=\frac{(A)}{N}\times\frac{(B)}{N}\times N$$

$$=\frac{(A)\times(B)}{N}$$

CRITERION OF INDEPENDENCE

When actual observation is equal to expectation, attributes are **independent** and there is no association between them. In such a case we expect to find the same proportion of A's amongst the B's as amongst the non-B's

Let us take an example :

Example 4. If (A) = people vaccinated = 50,
(B) = people not attacked by small-pox = 60,
(AB) = people vaccinated but not attacked = 20,
N = total number of people = 150,

it is required to find whether the attributes A (vaccination) and B (freedom from attack) are independent.

In this case, expectation of $(AB)=\frac{(A)\times(B)}{N}$

$$=\frac{50\times 60}{150}=20.$$

The actual observation (people vaccinated but not attacked) is thus equal to the expectation. Therefore, A and B are independent. We conclude that vaccination and freedom from attack are not associated with each other in the given case.

Let us take another example :

Example 5. If in the above example (AB) are not given, but instead, we know the number of people not vaccinated but attacked by small pox, *i.e.*, (ab), to be equal to 60, we proceed to find whether the attributes a and b are independent.

Expectation of $(ab)=\frac{(A)\times(B)}{N}$

Now $(a)=N-(A)=150-50=100$

and $(b)=N-(B)=150-60=90$

Therefore, $\frac{(a)\times(b)}{N}=\frac{100\times 90}{150}=60$

Again, the actual observation is equal to the expectation. In this case also the two attributes, a and b, are independent.

We can now put down the criterion of independence in more convenient form when actual class-frequencies of the second order only are given.

Attributes A and B are independent if,

Actual observation=expectation

that is, if $AB=\frac{(A)\times(B)}{N}$

Attributes a and b are independent if,

$$(ab)=\frac{(a)\times(b)}{N}$$

Attributes A and b are independent if,

$$(Ab)=\frac{(A)\times(b)}{N}$$

Attributes a and B are independent if,

$$(aB)=\frac{(a)\times(B)}{N}$$

$$\text{Hence, } (AB)\times(ab)=\frac{(A)\times(B)}{N}\times\frac{(a)\times(b)}{N}$$

$$=\frac{(A)\times(b)}{N}\times\frac{(a)\times(B)}{N}$$

$$\text{But, } \frac{(A)\times(b)}{N}\times\frac{(a)\times(B)}{N}=(Ab)\times(aB)$$

Therefore, $(AB)\times(ab)=(Ab)\times(aB)$

This last equation gives the required criterion of independence for the attributes A and B in the case of actual ultimate frequencies of the second order. We take one more example.

Example 6, Let the actual observations be as follows:—

People vaccinated but not attacked by small-pox $=60=(AB)$.
People not vaccinated and attacked by small-pox $=272=(ab)$.
People vaccinated but attacked by small-pox $=80=(Ab)$.
People not vaccinated and not attacked by small-pox $=204=(aB)$.

It is required to find whether the attributes A (vaccination) and B (freedom from attack by small-pox) are independent.

According to the criterion just indicated, A and B are independent only when

$$(AB)\times(ab)=(Ab)\times(aB)$$

Now, in the given case,

$$(AB)\times(ab)=60\times272=16{,}320$$
$$(Ab)\times(aB)=80\times204=16{,}320$$
$$\therefore (AB)\times(ab)=(Ab)\times(aB)$$

The criterion is satisfied and, therefore, the attributes A and B are independent, that is, not associated with each other.

Given A's and B's in a certain number of observations N and the fact that A and B are independent, we can calculate the values of second order frequencies.

Example 7. If there are 36 A's and 96 B's in 256 observations, how many AB's will there be, A and B being independent ?

$$\text{Now, } (AB) = \frac{(A) \times (B)}{N} \text{ if A and B are independent}$$
$$= \frac{36 \times 96}{256}$$
$$= 13.5$$

Example 8. In example 7 above, how many *ab's* will there be, *a* and *b* being independent ?

$$\text{Now, } a = N - (A) = 256 - 36 = 220$$
$$b = N - (B) = 256 - 96 = 160$$
$$(ab) = \frac{(a) \times (b)}{N} \text{ if } a \text{ and } b \text{ are independent}$$
$$= \frac{220 \times 160}{256}$$
$$= 137.5$$

ASSOCIATION AND DISASSOCIATION

In statistics the word association has a technical meaning, distinct from the one current in ordinary speech. Ordinarily one speaks of A and B as being associated if they appear together in a number of cases. It is not so in statistics, where A and B will be said to be associated only if they appear together in a larger number of cases than is to be expected if they are independent. The mere fact that *some* A's are B's, however great the proportion, is not enough to show that A and B are associated. This is a fundamental principle.

Association may be positive or negative. There is a simple way of knowing it. If two attributes, A and B, are not independent, but related to each other, then if

$$(AB) > \frac{(A) \times (B)}{N},$$

A and B are said to be **positively associated.** If, on the contrary,

$$(AB) < \frac{(A) \times (B)}{N}$$

A and B are said to be negatively associated, or, briefly, **disassociated.** It should be carefully noted that disassociation does not mean the same thing as independence.

In example 6, with the data as given, A and B are independent ; that is, they are not related. If the actual observation of people who were vaccinated but not attacked by small-pox, that is, if class-frequency (AB) were more than 60, and the expectation were 60, the attributes would have been positively associated. But if the actual cases of (AB) were less than 60, while expectation remained 60, A and B would have been disassociated.

In the light of the above principle, we take an example.

Example 9. Let $(A)=40$; $(B)=35$; $(a)=10$; $(b)=15$; $(AB)=30$; $(Ab)=10$; $(ab)=5$; $(aB)=5$; and, $N=50$.

Find the association between A and B, A and *b*, and *a* and B.

From the given observed frequencies we can construct a table like the following:

Table X (Observation)

	A	*a*	N
B	30	5	35
b	10	5	15
N	40	10	50

We now find the expectations.

$$\text{Expectation of } (AB) = \frac{(A) \times (B)}{N} = \frac{40 \times 35}{50} = 28,$$

Expectation of $(ab) = \frac{(a) \times (b)}{N} = \frac{10 \times 15}{50} = 3$,

Expectation of $(B) = \frac{35}{50} \times 50 = 35$,

Expectation of $(A) = \frac{40}{50} \times 50 = 40$, and so on.

We may now construct a table of expectations as below :

Table Y (Expectation)

	A	*a*	N
B	28	7	35
b	12	3	15
N	40	10	50

If we compare table X with table Y we shall be able to study the fact of positive and negative associations. In table Y, (AB) is 28, while in table X it is 30. This implies that actual observation of (AB) is greater than its expectation, or in other words,

$$(AB) > \frac{(A) \times (B)}{N}$$

Therefore, vaccination and freedom from attack, to make A and B stand for what they did so far in our examples, are positively associated.

On the other hand, (A*b*) in table Y is 12 and in table X it is 10. That is, actual observation of (A*b*) is less than its expectation. Therefore, A and *b* are negatively associated.

In table Y, *a*B is 7, while it is only 5 in table X. This implies that

$$(aB) < \frac{(a) \times (B)}{N}$$

Therefore, *a* and B are disassociated or negatively associated.

The Chi-Square Test.

An important contribution of Karl Pearson to statistical methodology has been the determination of the form of the distribution of Chi-Square and the development of methods of utilisation of this distribution. The quantity Chi-Square (represented by the symbol χ^2) is a measure of the degree to which observed frequencies deviate from the corresponding expected (or theoretical, or hypothetical) frequencies. In other words, Chi-Square measures the magnitude of the discrepancy between expectation (theory) and observation. Pearson determined the nature of the distribution of the values of χ^2 that would be obtained under given sampling conditions. Information about this distribution enables us to determine whether a given discrepancy between theory and observation may be attributed to chance, or whether it is the result of the inadequacy of the theory to fit the observed facts. Hence the importance of the Chi-Square test in statistical analysis.

An essential feature of the Chi-Square test is the comparison of the frequencies actually observed with the frequencies which upon some hypothesis are expected. Let m be the expected frequency, $(m+x)$ the observed frequency in any class and S be the sign for summation. Then, according to Fisher,

$$\chi^2 = \mathrm{S}\left(\frac{x^2}{m}\right),$$

the summation extending over all the classes. This formula gives the value of χ^2. A close glance at the formula would show that the smaller the difference between the observed frequency $(m+x)$ and the expected frequency (m), the smaller will be the value of χ^2. If observed frequency were equal to expected frequency, (which means, that if the attributes were independent), the value of χ^2 will obviously be zero.

The actual application of the Chi-Square test is facilitated by prepared tables. One such table is given by Fisher in his *Statistical Methods for Research Workers* (Table III, pp. 112-113 of 1948 edition). This table gives the values of Chi-Square for different levels of significance (from 0.01 to 0.99, that is, from 1% to 99%) for various values of n, n being the number of degrees of freedom in which the observation

may differ from the expectation. In order to use the χ^2 Table it is necessary to know the value of n. n is equal to the number of those classes whose frequencies can be filled arbitrarily.

The detailed procedure involved in using the Chi-Square Test for determining the fact of independence or association between two attributes may be illustrated with the help of an example.

Example 10. An investigation was carried out to determine whether there was any association between the eye colour of parents and eye colour of children. The eye colours were noted in the case of a random sample of 1,000 fathers and their eldest sons. In 471 cases both fathers and sons had light eyes, in 230 cases both had dark eyes, in 148 cases the fathers were dark-eyed and sons light-eyed, and in all remaining cases the sons were dark-eyed and the fathers were light-eyed. Required to determine whether eye colour in fathers and eye colour in sons is associated or independent.

From the given data we first prepare a table of observed frequencies as in Table X below :

Table X. *Observation.*

	Light-eyed Sons	Dark-eyed Sons	Total
Light-eyed Fathers	471	151	622
Dark-eyed Fathers	148	230	378
Total	619	381	1000

With the observed frequencies given in the four cells we may compare a set of expected frequencies, the expected frequencies being derived from the hypothesis that the eye colour in fathers has no relation with the eye colour in sons. That is,

the eye colour in fathers and the eye colour in sons are independent.

It will be seen from Table X that the marginal totals are in the ratio of 622: 378. On the hypothesis just mentioned, we proceed to calculate the expected frequencies. The 622 light-eyed fathers constitute 62.2% of all fathers, while the 378 dark-eyed fathers constitute 37.8% of the total. If the eye colour in fathers has no relation to the eye colour of sons, that is, if the two eye colours are independent, we should expect the 619 light-eyed sons to be divided between light-eyed fathers and dark-eyed fathers in this same proportion of 62.2% and 37.8%. In other words, 62.2% of the 619 light-eyed sons (that is, 385) should belong to the light-eyed fathers, while 37.8% of the 619 light-eyed sons (that is, 234) should belong to the dark-eyed fathers. The 381 dark-eyed sons should also be divided between the light-eyed fathers and dark-eyed fathers in exactly the same way, *viz.*, 62.2% (or, 237) to the light-eyed fathers and 37.8% (or, 144) to the dark-eyed fathers. (For convenience of calculation and for quicker understanding, fractions have been omitted. In fact, 385 should have been put down as 385.018, and other expected frequencies should also have been worked out to three places after the decimal point). The expected frequencies, so calculated, are given in Table Y below:—

Table Y. *Expectation.*

	Light-eyed Sons	Dark-eyed Fathers	Total
Light-eyed Fathers	385	237	622
Dark-eyed Fathers	234	144	378
Total	619	381	1000

These expected frequencies now conform to our hypothesis that eye colour in fathers and sons is independent. It will be seen that the marginal totals in Table X and the marginal totals in Table Y are identical. This means that the expected frequencies have been so calculated that the four frequencies are in proportion. These expected frequencies can also be calculated from the marginal totals in Table X, such that the expected frequencies should agree with the observed frequencies in the margins. If this procedure is adopted, the frequency of one cell alone need be calculated. For example, the expected frequency of light-eyed fathers that had light-eyed sons may be calculated as follows :

$$\frac{619 \times 622}{1000} = 385.018 = 385 \text{ approx.}$$

Having known this one frequency, the other three frequencies can be written down at once by subtracting from the marginal totals. Given 385, the expected frequency of light-eyed fathers that had dark-eyed sons is equal to $(622-385)=237$, and the expected frequency of dark-eyed fathers who had light-eyed sons is equal to $(619-385)=234$. The filling of these three cells determines the frequency of the fourth cell.

Now we should know the value of n. As already said, n is equal to the number of degrees of freedom in which observed frequencies may differ from the expected frequencies. Since we have found out only one expected frequency (*viz.*, 385) and have calculated the remaining three frequencies by more subtraction from the marginal totals, it is obvious that n is equal to 1, that is the observed frequencies can differ from the expected frequencies in only 1 degree of freedom. It should, therefore, be remembered that *in testing independence between two attributes in a* 2×2 *table like Table X, the observed frequencies can differ from the expected frequencies in only* 1 *degree of freedom.*

We may next calculate the value of Chi-Square. It was said that if m was the expected frequency and $(m+x)$ the observed frequency in any class, Chi-Square would be calculated thus :

$$\chi^2 = S\left(\frac{x^2}{m}\right)$$

This formula of Fisher can be put down slightly differently, as below, without in any way changing its meaning or effect :

$$\chi^2=\Sigma\left\{\frac{(f_0-f)^2}{f}\right\},$$

where, f_0 stands for the observed frequency in Table X and f for the expected frequency in Table Y. The X^2 in our example is calculated below, in this manner:—

Table Z. *Calculation of Chi-Square.*

f_0	f	f_0-f	$(f_0-f)^2$	$\frac{(f_0-f)^2}{f}$
471	385	+86	7396	19.2
148	234	−86	7396	31.6
151	237	−86	7396	31.2
230	144	+86	7396	51.3
1000	1000			133.3

$$\chi^2=\Sigma\left\{\frac{(f_0-f)^2}{f}\right\}$$
$$=133.3$$

The Chi-Square works out to be as high as 133.3. It will be recalled that we had started with the hypothesis that the eye colour in fathers and sons was independent ; if that were actually the case, the true value of Chi-Square should have been zero. It is, on the other hand, 133.3. Apparently, our hypothesis is wrong ; eye colour in fathers and sons is not independent but associated. But, how to be sure of this ? Can not this observed result be due to chance ? To test this we turn to the Chi-Square table. From this table, various values of Chi-Square for n, the number of degrees of freedom, equal to 1 are given below :

		Value of Chi-Square
For 1%, or 1 time out of 100	..	6.635
For 2%, or 2 times out of 100	..	5.412
For 5%, or 5 times out of 100	..	3.841
For 10%, or 10 times out of 100	..	2.706
etc.		

The entries in the above Chi-Square values for 1 degree of freedom indicate that if the true value were zero, a value as great as 3.841 would occur 5 times out of 100, as a result of chance fluctuations; and, a value as great as 6.635 would occur only 1 time out of 100 so that a value higher than 6,635 would not occur even once out of 100. The value of Chi-Square in our example, on the other hand, is 133.3. The probability of such a high value occuring is far less than 1 time out of 100. The observed result is, therefore, not consistent with the hypothesis with which we started. In other words, the discrepancy between the observed frequencies and expected frequencies is too great to be attributed to chance. Eye colour in fathers and eye colour in sons are, therefore, not independent but associated. That is, light-eyed fathers have a tendency to give birth proportionately more to light-eyed sons and the dark-eyed fathers proportionately more to dark-eyed sons.

It will be clear from the above illustration that in the Chi-Square Test we start with a null hypothesis; we start with the hypothesis that the given attributes are independent. If the value of Chi-Square turns out to be zero, our hypothesis is apparently established; if not, we have to examine with the help of the Chi-Square table as to whether the differences between the observed and expected frequencies could be due to chance. If it is found that they could not be due to chance, our hypothesis falls; the attributes are not independent but associated.

Independence, Association & Disassociation.

The mere fact that some B's are A does not mean that A and B are associated; we must also know the proportion of *b* that are A in order to determine whether A and B are indepen-

dent, associated or disassociated. Consider the three tables 1, 2 and 3 below:

Table 1

	A	*a*	N
B	80	20	100
b	80	20	100
N	160	40	200

Table 2

	A	*a*	N
B	80	20	100
b	60	40	100
N	140	60	200

Table 3

	A	*a*	N
B	80	20	100
b	90	10	100
N	170	30	200

In all the three tables 80% of B's are A. But whereas in table 1 80% of *b*'s are also A, in table 2 only 60% of *b*'s are A and in table 3,90%. We find that A and B are independent in table 1, associated in table 2 and disassociated in table 3, whatever the criterion of judging the fact of independence that we might follow.

It will be seen that table 1 is at the base of the Chi-Square test discussed above; the value of χ^2 in its case will work out to be zero. Tables 2 and 3 show that the χ^2 will not be zero.

CO-EFFICIENT OF ASSOCIATION

So far we have ascertained the fact of association by comparing the observed class-frequencies with the expectations. We have not measured the degree of association. Several co-efficients have been devised for judging the degree or intensity of association. Of these, the following co-efficient due to Yule is the simplest:

$$Q=\frac{(AB)(ab)-(Ab)(aB)}{(AB)(ab)+(Ab)(aB)}$$

where, Q stands for the Co-efficient of Association. This Co-efficient is zero when the attributes are independent, +1 if they are completely associated, and −1 if they are completely disassociated.

Let us take an example.

Example 11. Required to compute the co-efficient of association between A and B from the data given in table X of example 9, above.

$$Q=\frac{(AB)(ab)-(Ab)(aB)}{(AB)(ab)+(Ab)(aB)}$$

$$=\frac{(30\times5)-(10\times5)}{(30\times5)+(10\times5)}$$

$$=\frac{100}{200}$$

$$=+0.5$$

Hence the intensity of association between the attributes A and B is 0.5 and the association is positive.

Let us take another example.

Example 12. Required to compute the co-efficient of association between A and B from the data given in example 1, above.

$$Q=\frac{(AB)(ab)-(Ab)(aB)}{(AB)(ab)+(Ab)(aB)}$$

$$=\frac{(100\times40)-(50\times80)}{(100\times40)+(50\times80)}$$

$$=0.$$

Hence the attributes A and B are independent.

So far we have measured the intensity of association between A and B. Let us take an example to find out the degree of association between a and B.

Example 13. Required to compute the coefficient of association between a and B from the data given in example 11, above.

The formula for coefficient of association between A and B is,

$$Q=\frac{(AB)(ab)-(Ab)(aB)}{(AB)(ab)+(Ab)(aB)}$$

To convert this formula into one for association between a and B, change A wherever it occurs in the formula into a and a into A. Thus, the formula for coefficient of association between a and B would be,

$$Q=\frac{(aB)(Ab)-(ab)(AB)}{(aB)(Ab)+(ab)(AB)}$$

$$=\frac{(5\times10)-(5\times30)}{(5\times10)+(5\times30)}$$

$$=-\frac{100}{200}$$

$$=-0.5$$

Another way of finding the intensity of association between a and B from the data given in example 11 would be to retain in tact the formula for association between A and B, and change A into a and a into A in the class-frequencies given in example 11. Accordingly, (AB)=5, (ab)=10, (Ab)=5, (aB)=30. Then,

$$Q=\frac{(AB)(ab)-(Ab)(aB)}{(AB)(ab)+(Ab)(aB)}$$

$$=\frac{(5\times10)-(5\times30)}{(5\times10)+(5\times30)}$$

$$=-0.5$$

Hence the intensity of association between the attributes a and B is 0.5 and the association is negative.

In like manner, degree of association or disassociation between any two attributes can be estimated.

Comparison of intensity of Association.

The Coefficient of Association can also be utilised to compare the intensity of association between two attributes with the intensity of association between two other attributes.

Example 14. From the figures given in the following table, compare the association between literacy and unemployment among males in rural and urban areas:

	Urban (Lakhs)	Rural (Lakhs)
Total males (N)	25	200
Literate males (A) ..	10	40
Unemployed males (B) ..	5	12
Literate and unemployed males (AB)	3	4

We have first to find out the second-order frequencies of the urban and rural areas. For the urban areas:

$$(A)=(Ab)+(AB) \qquad (B)=(aB)+(AB)$$
$$10=(Ab)+3 \qquad 5=(aB)+3$$
$$\therefore (Ab)=7 \qquad \therefore (aB)=2$$

$$(a)=(aB)+(ab)$$
$$15=2+(ab)$$
$$\therefore (ab)=13$$

Similarly for the rural areas $(Ab)=36$; $(aB)=8$; $(ab)=152$.

Now, coefficient of association between A and B in the rural areas may be calculated thus:

$$Q=\frac{(AB)(ab)-(Ab)(aB)}{(AB)(ab)+(Ab)(aB)}$$
$$=\frac{(3\times13)-(7\times2)}{(3\times13)+(7\times2)}$$
$$=+0.47$$

And, coefficient of association between A and B in rural areas:

$$Q = \frac{(AB)(ab) - (Ab)(aB)}{(AB)(ab) + (Ab)(aB)}$$

$$= \frac{(4 \times 152) - (36 \times 8)}{(4 \times 152) + (36 \times 8)}$$

$$= +0.36$$

Thus the intensity of association between literacy and unemployment among males is greater in urban areas than in rural areas.

Yule's co-efficient of association is quite easy to compute and is a convenient measure of association, since it not only exhibits the intensity of positive and negative associations, but also shows the independent character of the attributes.

PARTIAL ASSOCIATION

If in a given case it is found that

$$(AB) > \text{or} < \frac{(A) \times (B)}{N}$$

all that this information leads us to is that A and B are related with each other in some way. We cannot say whether the relationship is direct or of any other kind. It is possible that association between A and B may not be direct, but due to the association of A with C and of B with C. An example will make the point clear.

An association is observed between 'vaccination' and 'exemption from attack by small-pox', that is, more of the vaccinated people are exempt from attack than the unvaccinated ones. It may be argued that this does not imply that vaccination protects the people from attack, but that most of the unvaccinated people are drawn from those living in insanitary and filthy conditions. Thus A (vaccination) and B (exemption from attack) are associated due to the association of both with C (hygienic conditions).

The ambiguity in the above case arises from the fact that the universe contains not only the objects possessing the third attribute alone, or objects not possessing it, but both. In our example, both hygienic and non-hygienic conditions may be prevailing in the locality where observations have been made.

If, however, the universe of observation were confined to either class alone (for instance, the observations relating to vaccination and attack were made from a narrow section of the population living under approximately identical hygienic conditions), and still A and B were found to be associated, the above ambiguity would not arise.

The associations found between the attributes A and B in the universe of C's and the universe of *c*'s are termed as **partial** associations, to distinguish them from total associations found between A and B in the universe at large.

Partial association may prove to be entirely misleading, for what is true of the whole is not true of each of the parts. To return to our example, observations regarding vaccination and attack may be drawn from people living under same hygienic conditions, yet some of the people may be rich and others poor. There may be a positive association between vaccination and exemption from attack among the rich but not among the poor. The disparity between these results may be explained by the simple fact that the poor are more open to attack than the rich, so that attack is not independent of poverty. Or, it may be that only the rich get themselves vaccinated, and vaccination is, in this case, not independent of poverty.

Thus an illusory or misleading association may arise in a case where in the given universe there exists a third attribute C with which both A and B are associated, positively or negatively. If both the associations are of the same sign, the resulting illusory association between A and B will be positive; if of opposite signs, the illusory association will be negative. For example, if the associations between A and C, and between B and C are positive, they would give rise to an illusory positive association between A and B.

Illusory association may also arise in a different manner, that is, through the personality of the observer. If the attention of the observer fluctuates, it is likely that he may observe the presence of A when he observes the presence of B, and *vice versa*. In such a case, A and B will both be associated with the observer's attention C, and an illusory association will result.

CONTINGENCY

Manifold classification.

We have so far dealt with cases in which the universe was divided into two parts by a simple dichotomy. A universe may also be divided into a number of parts by a similar process. For example, suppose there is a universe of school-boys. Following the classification by dichotomy, we can divide this universe into "those with black hair" and "those not with black hair". We can extend this dichotomous classification to a three-fold division: "those with black hair", "those with grey hair", and "those with neither black nor grey hair". We may further extend this three-fold division to a four-fold division by adding the category of "those with red hair"; and so on. Any classification in which we make more than two divisions is called Manifold classification.

Contingency table.

Tables X and Y given in the foregoing pages are 2×2 tables presenting a dichotomous classification. A Contingency Table is a table which presents the different class-frequencies in respect of a manifold classification. In a three-fold classification our universe may be divided first according to s heads, A_1, A_2, $A_3 \ldots\ldots A_s$. Each of the classes so obtained may then be divided into t heads, B_1, B_2, $B_3, \ldots\ldots B_t$. Then there will be $s \times t$ ultimate classes, whose frequencies will be put down as in the specimen table given below:

Contingency Table.

		Attribute A						
		A_1	A_2	..	..	A_{s-1}	A_s	Totals
Attribute B	B_1	(A_1B_1)	(A_2B_1)	..	..	$(A_{s-1}B_1)$	(A_sB_1)	(B_1)
	B_2	(A_1B_2)	(A_2B_2)	..	..	$(A_{s-1}B_2)$	(A_sB_2)	(B_2)
	..	..	..	..	..	..	..	..
	B_t	(A_1B_t)	(A_2B_t)	..	..	$(A_{s-1}B_t)$	(A_sB_t)	B_t
	Totals	(A_1)	(A_2)	..	..	(A_{s-1})	(A_s)	N

The above table is called Contingency Table. In this table, the frequency of the class A_2 B_2 is entered in the cell common to the A_2 column and the B_2 row. The totals at the end of each row and those at the feet of each column, give the first order frequencies; the grand total of all frequencies in the bottom right-hand corner gives the total number of observations, N. It should now be clear that contingency table results from manifold classification according to two attributes, A and B.

Association in Contingency Tables.

Association between A's and B's can be studied in a contingency table by reducing the table to a 2×2 table (or nine-square table like table X in the foregoing pages) drawn on the basis of dichotomous classification. The procedure thereafter is the same as followed in several examples already.

For reducing a contingency table to a 2×2 table, the criterion would be to pool together only the like columns, and similarly the like rows. Suppose attribute A, hair-colour, is classified into fair, brown, black and red and attribute B, eye-colour, is divided into blue, grey and brown in a contingency table. Then, fair, brown and black hair-colours may be lumped together so that this pooled class will be called 'not red'. Similarly, blue and grey eye-colours may be pooled together to be called 'not brown' eye-colour. Association between A and B may then be measured in the usual manner.

Coefficient of Contingency.

The above process is, however, laborious, particularly if the number of classes into which A's and B's are divided is large. Further, the question before us may not be only to know the association between A and B, but also to find whether A's *on the whole* are distinctly dependent on B's, and if so, whether the dependence is very close or revers, and what its intensity is. For this purpose, we have to use the Coefficient of Contingency. This coefficient gives us the nature and extent of dependence in a manifold classification, just as the coefficient of association gives us the nature and intensity of association in a dichotomous classification.

EXERCISES

(1) How would you distinguish between 'association' and 'correlation' as the terms are used in statistics?

(M.A., Agra, 1944).
(M. Com., Alld., 1948).

(2) Given the following ultimate class-frequencies, find the frequencies of the positive and negative classes and the whole number of observations, N:—

$(AB)=200$; $(Ab)=100$
$(aB)=160$; $(ab)=80$

(3) From the following data find whether the attributes A and B are independent:

$(A)=100$, $(B)=120$, $(AB)=40$, $N=300$.

(4) If $(AB)=120$, $(ab)=544$, $(Ab)=160$ and $(aB)=408$, find whether the attributes A and B are independent.

(5) Given the following ultimate class-frequencies, find the frequencies of the positive classes.

$(ABC)=298$ $(AbC)=450$ $(aBC)=408$ $(abC)=342$
$(ABc)=1476$ $(Abc)=2292$ $(aBc)=3524$ $(abc)=43684$

(6) Given the following frequencies of the positive classes, find the frequencies of the ultimate classes:

$(N)=47,426$ $(ABC)=312$
$(A)=3,236$ $(AB)=856$
$(B)=4,030$ $(AC)=670$
$(C)=1,540$ $(BC)=312$

(7) Show whether A and B are independent, positively associated or negatively associated in the following cases:—

(1) $N=1000$ $(A)=470$ $(B)=620$ $(AB)=320$
(2) $(AB)=512$ $(aB)=1536$ $(Ab)=96$ $(ab)=288$
(3) $(A)=245$ $(AB)=147$ $(a)=285$ $(aB)=190$

(8) Find the association between eye-colour of husband and eye-colour of wife from the following data:—

Husbands with light eyes and wives with light eyes ..	1236
Husbands with light eyes and wives with not light eyes ..	856
Husbands with not light eyes and wives with light eyes ..	528
Husbands with not light eyes and wives with not light eyes	476

(9) Can vaccination be regarded as a preventive measure for small-pox from the data given below?

'Of 1,482 persons in a locality exposed to small-pox, 368 in all were attacked'.

'Of 1,482 persons, 343 had been vaccinated and of these only 35 were attacked'.

(M. Com., Alld., 1944).

(10) The following table gives the number of persons suffering from certain infirmities in Bengal in 1931:—

Sex	Total Number	Insane	Deaf-mutes	Deaf-mutes and Insane
Males	260 lakhs	12,650	21,301	545
Females	241 ,,	9,055	14,136	317

Trace the association between insanity and deaf-muteness for males and females of Bengal separately.

(M.A., Alld., 1938).

(11) Find out the coefficient of association between the type of college training and success in teaching from the following table:—

Institution		Successful	Unsuccessful	Total
Teacher's College	..	58	42	100
University	..	49	51	100
Total	..	107	93	200

(M.A., Alld., 1950).

(12) What do you understand by contingency?

Calculate the coefficient of association between the race of

relief receipients and type of relief from the following contingency table:—

Type of relief	Race		Total
	Negro	White	
Work ..	72	376	448
Direct ..	154	250	404
Total ..	226	626	852

(M.A., Alld., 1951).

(13) Explain 'Association' and advantages of studying it.

A census revealed the following figures of the blind and insane in two age groups in a certain population:

	Age group 15—25 years	Age group over 25 years
Total population ..	2,71,000	1,60,200
Number of blind ..	1,000	2,000
Number of insane ..	600	1,000
Number of insane among the blind	19	9

(*a*) Obtain a measure of association between blindness and insanity in each of the two age groups.

(*b*) Are blindness and insanity associated or disassociated ?

(M. Com., Alld., 1950).

(14) From the following table, test whether the colour of son's eye is associated with that of father's.

		Eye colour in sons	
		Not Light	Light
Eye colour in fathers.	Not Light	230	150
	Light	150	470

(M. Com., Alld., 1952).

(15) Prove that the following data are inconsistent:—

N=1,000	(A*b*)=483
(A)= 525	(A*c*)=378
(B)= 312	(B*c*)=226
(C)= 470	(ABC)= 25

(M. Com., Alld., 1952).

(16) (a) If there are 144 A's and 384 B's in 1024 observations, how many AB's will there be, A and B being independent?

(*b*) In the above, what would be the number of *ab*, if *a* and *b* are independent?

(17) The Fortune magazine of the U.S.A. published the following results of a sample survey of public opinion regarding election of Roosevelt as the president of U.S.A.

Attitude towards election	Rich	Poor	Total
Favourable ..	508	1559	2067
Unfavourable ..	905	1114	2019
Total ..	1413	2673	4086

Is attitude towards election issue guided by the economic status of the voters? What test would you apply? The following table of 1% values of Chi-Square is reproduced for your use:—

Degree of freedom.	1	2	3	4
1% value of Chi-Square	6.635	9.210	11.341	13.277

(M.A., Patna, 1944).

(18) Given: (A)=10; (*a*)=20; (*B*)=12; (*b*)=18 and (AB)=4. Find out whether the attributes A and B, *a* and *b*, A and *b*, and *a* and B are independent, associated or disassociated.

(19) The results of a survey in an Indian city regarding the habit of tea taking are given below:

Number of	Hindus	Muslims
Families taking tea	636	76
Families not taking tea	301	20

Discuss whether there is any significant difference between the two communities in the matter of taking tea. You may use the following values of Chi-Square:

Degrees of Freedom (n)	1	2	3
1% values of Chi-Square ..	6.64	9.21	11.34
5% values of Chi-Square ..	3.84	5.99	7.81

(20) In the course of anti-malarial work in Birnagore in the third quarter of 1932, quinine was administered to 606 adults out of a total population of 3,540. The incidence of malarial fever is shown below. Discuss the preventive value of quinine.

	Fever	No fever	Total
Quinine	19	587	606
No Quinine	193	2741	2934
Total	212	3328	3540

You may use the 5% value of Chi-Square for n, degrees of freedom, equal to 1, the value being 3.841.

(M.A., Cal., 1937).

(21) The following is a summary of the statistical features of a census of ration cards:—

Item No.	Category		Total number of Cards belonging to the Category
1	The whole of the Census.	..	10,000
2	Permanent residents.	..	5,100
3	Males.	..	4,900
4	Consumers of rice.	..	4,270
5	Permanent male residents.	..	1,890
6	Consumers of rice among permanent residents.	..	1,400
7	Males consuming rice.	..	970

Show that the entry against item No. 7 is inconsistent with the entries against all the previous items, namely 1, 2, 3, 4, 5 and 6 taken together.

(M. Com., Alld., 1951).

(22) The following table shows the distribution of temper in pairs of sisters in an exhaustive school enquiry:—

	First sister		
Second sister	Good natured	Sullen	Total
Good natured ..	1040	180	1220
Sullen ..	160	120	280
Total ..	1200	300	1500

Trace the association, if any, in the distribution of tempers in first sisters and second sisters.

(M. Com., Raj., 1952).

(23) The male population of the U. P. is 250 lakhs. The number of literate males is 20 lakhs, and total number of male criminals is 26 thousand. The number of literate male criminals is 2 thousand. Do you find any association between literacy and criminality?

(M.A., Agra, 1943).

(24) Investigate whether there is any association between extravagance in father and son from the following data:—

Extravagant sons with extravagant fathers	..	496
Miser sons with extravagant fathers	..	162
Extravagant sons with miser fathers	..	184
Miser sons with miser fathers	..	1158

(25) Write explanatory notes on the following:

Association; Criterion of Independence; Partial Association; Disassociation; Chi-Square Test; Contingency.

(26) Criticize the following arguments:—

(1) 99 per cent of the people who take alcohol die before they reach the age of 80 years. Therefore, taking alcohol is bad for longivity.

(2) 99 per cent of the members who voted for the tenancy bill were cultivators. Therefore, it was unfair to suppose that the voting was unbiassed.

(27) Explain with the help of an example as to how you would use the Chi-Square test for studying the fact of independence or association between two attributes.

CHAPTER XVIII

INTERPOLATION

Meaning of Interpolation.

Interpolation means framing the most likely estimate of a certain quantity under certain assumptions. In Chapter IX, the mode and the median were interpolated in the modal and the median classes respectively. This was done by proceeding with certain assumptions in both the cases. In locating the mode in a continuous frequency distribution it was assumed that mode was influenced by the class-intervals adjacent to the modal class ; while, in locating the median in a similar series it was assumed that the magnitude of the median class was uniformly distributed over its frequencies. Location of the mode and the median in a grouped distribution suggests examples of interpolation, as also the usefulness of this device for estimating some missing figure in a series.

Necessity of Interpolation.

It is not only to estimate the best values of mode and median in a continuous frequency distribution that the method of interpolation is employed. There are many other cases where gaps may have to be filled in. Such gaps may arise from the fact that no record has been made, or its details are insufficient, or it has been lost or destroyed. For example, it may be necessary for purposes of administration for a State or Central Government to know with a reasonable degree of accuracy the population of the State or the country for intercensal years. Similarly, a sociologist may like to know the number of people in different age-groups during the intercensal period, an economist may desire to have a knowledge of the total tax-revenue raised in a certain year, while a businessman may like to fill up the gaps in his yearly sales register. In all these cases it cannot be supposed that the figures relating to a particular past year would apply to the year whose figures are required to be estimated. Nor can mere imaginary figures be relied upon. The most likely estimate has to be made.

Such an estimate may relate to some past date or to future one. The technique of estimating a past figure is termed as **Interpolation** while that of estimating a probable figure for the

future is called **Extrapolation.** To make an estimate certain assumptions are necessary.

Assumptions.

The first assumptions that is made in interpolation or extrapolation is that there are no sudden jumps from one period to another. If population figures for India for 1911, 1921, and 1941 are given and an estimate has to be made for the figure for 1931, this would be done only when it is assumed that there was no violent disturbance in the intermediate dates, nor was the year 1931 an exceptional year such as that affected by epidemics, famine, war or other calamity, or large scale migration.

The second assumption is that in the absence of evidence to the contrary the rise or fall has been uniform. That is, in our example, the population growth has to be assumed to be uniform between 1921 and 1941, if the year 1911 or some other information has not to modify this assumption.

Accuracy of Interpolation.

Upon the above assumption figures may be interpolated, but the question that arises is: 'What is the certainty that the interpolated figures, which by hypothesis are unknown, are in reality the most probable figures?' In the words of Dr. Bowley, the accuracy of interpolation depends "(1) on knowledge of the possible fluctuations of the figures, to be obtained by a general inspection of the fluctuations at dates for which they are given; (2) on knowledge of the course of the events with which the figures are connected."[1]

It follows, therefore, that in basing arguments upon such figures the fact that they are interpolated ones should not be lost sight of. Interpolated figures are based on quite a different class of evidence from those which result from direct evidence or count. In some instances, interpolations may represent figures which do not exist and which are used only for convenience of calculation. For instance, in allotting monthly marks to a student who was absent from a few seminars, attention may be paid to the student's general place in the class and to the average marks got by the students present in those particular seminars. Marks thus allotted have no exist-

[1] Bowley, A.L., *Elements of Statistics,* 1920 ed., page 217.

ence. In other cases, interpolated figures may be the most probable estimates of figures that really exist. Therefore, all such estimates must be indicated as interpolations; it is always better to point out the method used in obtaining them. If any subsidiary information, which may be regarded as a direct evidence of the accuracy of interpolated figures, is available, it will be well to state it also. Further, if practicable, interpolated figures should be stated not as exact ones, but as lying in a range within which their accuracy may not be questioned.

Methods of Interpolation.

Figures can be interpolated by the graphic method or by algebraic treatment. Graphic method is good to follow when the quantities show cyclical character. We shall discuss below, with suitable examples, the graphic method, the method of fitting a parabolic curve, the method of advancing differences (Newton's method) and the Lagrange's formula.

GRAPHIC METHOD

Graphic Method in a Continuous Time Series.

This method may be explained by an example. Table 64 gives the population of undivided Bengal during the seven censuses between 1881 and 1941. Column 1 of the table shows the independent variable (x) which advances by an equal increment of 10 years. Column 2 shows the corresponding values of the dependent variable (y).

Table 64. *Population of undivided Bengal.* 1881—1941.

Year	Population in lakhs
x	y
1881	363
1891	391
1901	421
1911	455
1921	567
1931	501
1941	603

Suppose the figure 421 for the year 1901 is not given, and we are required to interpolate it. The task of interpolation would depend upon the evidence available for the purpose. If, for example, we know only the figures 391 and 455 relating to the years 1891 and 1911 respectively, we may plot them on a graph paper with years on the base line and population on the vertical scale. Since only a straight line would result from joining the points, we have no alternative but to assume that the population between 1891 and 1911 rises at a uniform rate. In the absence of any information to the contrary this is the most correct assumption possible. The height of the ordinate drawn from 1901 to intersect the straight line would give us an estimate of population of undivided Bengal in 1901. This would be equal to $\frac{544+391}{2}$ or 423 lakhs, which exceeds the actual figure (421 lakhs) by only 2 lakhs. The difference is not very great, the error being of a little less than 0.5 per cent. If an estimate of population for any other year during the intercensal period is required, an ordinate from the particular year may be raised to intersect the straight line. The height of the ordinate on y would give the figure for the particular year.

Suppose figures are available for all the census years from 1881 to 1941, except 1901. The various y's in that case shall be plotted against their respective x's on a graph. If the resulting points are joined as straight lines with a ruler, we will have to assume sudden jumps in the growth of population, *cf.* points C, D, and E in Figure 33. But, it has been said above that rather than making such an assumption, we should assume that sudden changes in the quantities from one period to another do not occur. Therefore, instead of joining the points by straight lines, we should draw through them all a line whose curvature is as smooth as possible. Such a curve may be constructed on mathematical principles or drawn freehand. In Figure 33 it has been drawn freehand, connecting each (x, y) point. To find the y proper to 1901, we have drawn an ordinate through 1901 intersecting the curve at P. The height of this ordinate, PM, which is 421 gives the figure

in lakhs for the population of undivided Bengal in 1901. The remarkable closeness with which the interpolated figure agrees with the actual one given in table 64 is more or less accidental.

Interpolation of Population of undivided Bengal by Graphic Method

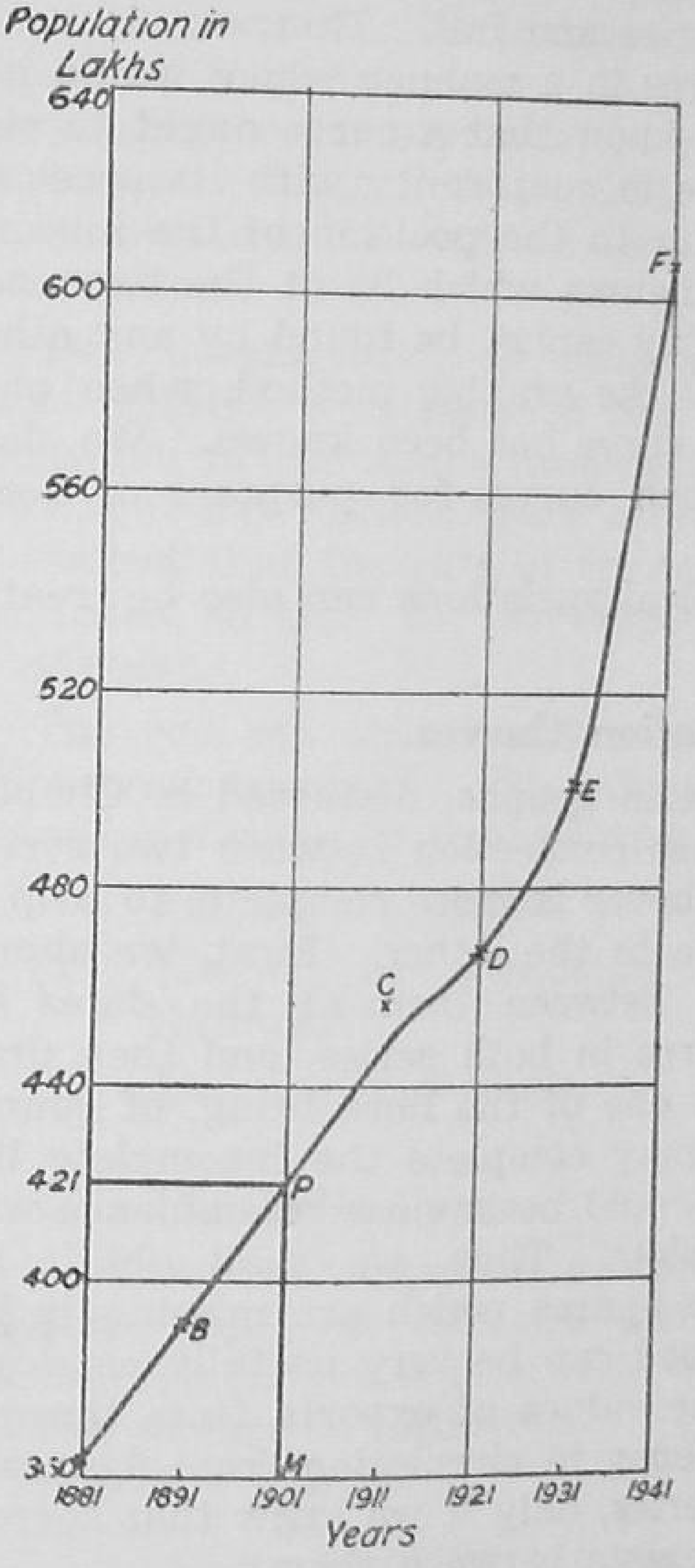


Fig. 33.

The principle followed in Figure 33 is not an unreasonable one to adopt, for, in effect, it gives due weight to each of the observations (y's) actually recorded, and it assumes an even course from each year to the next—a quite justifiable assumption in the absence of any evidence that some sudden discontinuity or break has taken place in the y's.

Graphic Method in Seasonal & Cyclical Series.

If we have a series of monthly data relating to a certain phenomenon, say, sales of silver or price of wheat in India, and the data show seasonal fluctuations, which we can study by the method discussed while dealing with Analysis of Time Series in Chapter XIV, we can interpolate figures for any month for which the records are incomplete. The amount of sales of silver in India would, for instance, show a rise during summer every year, since that is the marriage season in the

country; and, the price of wheat would show a fall in April-May when new crop arrives in the market. The curves drawn for such phenomena would exhibit a kind of regular variation, that is, they would regularly rise and fall. This would enable the filling in of unknown figures in a manner which would not be unsatisfactory. For, if we know that a curve ought to rise or fall to a certain limit to be in conformity with its seasonal variation, we get a reliable clue to the position of the missing figures. Further, even those figures which lie at the two ends of our series and which probably cannot be found by any other method, can be traced out by the graphic method, when once the nature of variation in the curve has been known. We shall see later the usefulness of such curves for purposes of forecasting.

Annual data showing cyclical variations can also be treated in a similar manner.

Graphic Method and Correlation Curves.

If, by the use of correlation graphs, discussed in Chapter XV, we are able to find a close connection between two series, we can use one of them, whichever is more complete, to help in interpolating a missing figure in the other. First, we should carefully study the closeness between them at the dates for which we have complete figures in both series, and then draw a figure similar to Figure 30, one of the lines being, of course, incomplete. Thereafter, we may complete the incomplete line such that the completed line would be in close resemblance with the line that is already complete. Thus, we shall obtain the most probable values for the figures which are missing in the incomplete series. This method can be very usefully employed in interpolating figures of the values of exports from those of imports, of the amount of money in circulation from figures of prices, and for many other series, only if we know that correlation of a very high degree exists between them.

ALGEBRAIC TREATMENT

The problem of interpolation to which greatest attention has been paid is as follows :

If one quantity is subject to continuous regular change, and a second quantity changes in connection with it, and if we

know or can estimate directly only some discontinuous value of this second quantity, then it is required to estimate the most probable value of the second quantity corresponding to given values of the first. For instance, given the annual premia payable on a life policy at ages 25, 30, etc. years, it is required to interpolate the premium for intermediate ages; or, given the population of India in 1881, 1891, 1901 and 1911 it is required to interpolate it for intermediate dates or extrapolate it for future dates.

Two assumptions, as already noted, are made in such cases. Firstly, it is assumed that the quantity (premium, or population, in the above examples) changes continuously, that is, without any sudden break at any figure. And, secondly it is assumed that the rate of change of the quantity is likewise continuous, so that the curve representing it is smooth, and not angular.

The problem stated above can be tackled systematically by algebraic treatment. We take up below three of the several methods available for interpolation:—

(1) Fitting a parabolic curve,

(2) Method of advancing differences, or, Newton's formula.

(3) Lagrange's formula.

Fitting a Parabolic Curve.

To make the argument as general as possible we shall speak of x and y as variables and assume the value of y as depending on that of x in such a way that when x is given, y is known or can be estimated.

Suppose

$$y=a+bx+cx^2+\dots\dots\dots\dots$$

where, a, b, c......are constants to be determined, and their number can be made to depend upon the number of known values of y. The equation

$$y=a+bx+cx^2+\dots\dots\dots nx^n$$

represents a curve called a parabola of the nth order.

Let us illustrate the method by fitting a parabolic curve to the following figures giving the population of Allahabad at decennial censuses:—

Year	1901	1911	1921	1931
Population in thousands	172	171.7	157.2	183.9

It is required to interpolate the population for 1916. Assuming that no abnormal conditions prevailed in 1916 to cause a sudden change in the population of Allahabad, let us proceed to estimate the population for that year with the help of the given data. Since the known points are four in this particular case, we would take as the curve through them a parabola of the 3rd order, *viz.*,

$$y = a + bx + cx^2 + dx^3 \ldots\ldots\ldots\ldots (1)$$

Then, the four known points would be just sufficient to determine the four constants a, b, c, d. Now, the x class-intervals are equal, being of 10 years each; we measure from 1916 as origin, and get

$x=$	-15,	-5,	o,	$+5$,	$+15$
$y=$	172,	171.7,	y_0,	157.2,	183.9

where y_0 is the number to be estimated.

To further simplify the algebra we may take 5 years as unit for x, so that

$x=$	-3,	-1,	o,	$+1$,	$+3$
$y=$	172,	171.7,	y_0,	157.2,	183.9

Since all five points are to lie on the curve with equation as in (1), we substitute the above values of x in the equation, and get

$$172 = a - 3b + 9c - 27d \quad \ldots\ldots\ldots\ldots \quad \text{(i)}$$
$$171.7 = a - b + c - d \quad \ldots\ldots\ldots\ldots \quad \text{(ii)}$$
$$y_0 = a \quad \ldots\ldots\ldots\ldots \quad \text{(iii)}$$
$$157.2 = a + b + c + d \quad \ldots\ldots\ldots\ldots \quad \text{(iv)}$$
$$183.9 = a + 3b + 9c + 27d \quad \ldots\ldots\ldots\ldots \quad \text{(v)}$$

Equation (iii) tells us that y_0, the value that we have to interpolate, is equal to the constant a. Therefore, from equations (i), (ii), (iv) and (v) we proceed to deduce the value of the constant a.

Adding (i) and (v),

$$2a+18c=172+183.9 \quad \ldots\ldots\ldots\ldots \quad \text{(vi)}$$

Adding (ii) and (iv),

$$2a+2c=171.7+157.2$$

$$\text{or } 18a+18c=9(171.7+157.2) \quad \ldots\ldots\ldots\ldots \quad \text{(vii)}$$

Subtracting (vi) from (vii),

$$16a=9\ (171.7+157.2)-(172+173.9)$$
$$=2960.1-355.9$$
$$=2604.2$$

Therefore, $a=162.7625$

or, $y_0=162{,}763$

The population of Allahabad, as interpolated, is 162,763 for 1916.

In a like manner, interpolation of population for any other intercensal year or extrapolation for any year after 1931 can be made.

It should be remembered that if data were given for three years, equation giving the parabola of the 2nd order, *viz.*, $y=a+bx+cx^2$, would have been used. Similarly, if data were given for five periods, the equation used would have been that of the 4th order; and so on.

The above method results in simultaneous equations, like the equations, (i), (ii), (iv) and (v), and is also sometimes called the method of simultaneous equations.

Newton's Method of Advancing Differences.

Table 65 gives in col. (2) the amount of annual premia required by an insurance company at different ages in col. (1) to secure Rs. 1,000 without profits. It is required to calculate the amount of premium payable at the age of 22 next birthday.

Table 65. *Annual Premia on a life policy of Rs. 1,000, and their differences.*

(1) Age next birth-day in years x	(2) Annual premium in Rs. y	Differences (3) First Δ	(4) Second Δ^2	(5) Third Δ^3	(6) Fourth Δ^4	(7) Fifth Δ^5
20 x_0	25 y_0					
		3 Δ_0				
25 x_1	28 y_1		1 Δ^2_0			
		4 Δ_1		0 Δ^3_0		
30 x_2	32 y_2		1 Δ^2_1		0.5 Δ^4_9	
		5 Δ_2		0.5 Δ^3_1		$-.25$ Δ^5_0
35 x_3	37 y_3		1.5 Δ^2_2		0.25 Δ^4_1	
		6.5 Δ_3		0.75 Δ^3_2		
40 x_4	43.5 y_4		2.25 Δ^2_3			
		8.75 Δ_4				
45 x_5	52.25 y_5					

Each entry in the difference columns is formed by taking the *algebraic* difference of the entries on the left. Thus,

(1) $\Delta_0 = y_1 - y_0 = 28-25=3$;
(2) $\Delta^2_0 = \Delta_1 - \Delta_0 = 4-3=1$;
(3) $\Delta^3_0 = \Delta^2_1 - \Delta^2_0 = 1-1=0$;
(4) $\Delta_1 = y_2 - y_1 = 32-28=4$;
(5) $\Delta^2_1 = \Delta_2 - \Delta_1 = 5-4=1$;
(6) $\Delta^5_0 = \Delta^4_1 - \Delta^4_0 = .25 - .5 = -.25$

In this manner, differences have been calculated in main columns 3, 4, 5, 6 and 7 of the table.

The formula to be used for interpolating the value of y for a given x, due to Newton, is,

$$y_x = y_0 + x\Delta_0 + \frac{x(x-1)}{1\times 2}\Delta^2_0 + \frac{x(x-1)(x-2)}{1\times 2\times 3}\Delta^3_0 + \frac{x(x-1)(x-2)(x-3)}{1\times 2\times 3\times 4}\Delta^4_0 \ldots\ldots$$

where, y_x is the figure to be interpolated, Δ's are the differences, and x is calculated as follows :

$$\frac{\text{Year of interpolation—Year of origin}}{\text{Time distance between adjoining years}}$$

The above equation uses differences up to the 4th order, and can be extended to the nth order.

It is required to know the amount of premium payable at age 22 years next birthday.

Now, from the table we find,

$$y_0 = 25 \; ; \; x = \frac{22-20}{25-20} = \frac{2}{5} = .4 \; ;$$

$$\Delta_0 = 3; \; \Delta^2_0 = 1 \; ; \; \Delta^3_0 = 0 \; ; \; \Delta^4_0 = .5 \; ; \; \Delta^5_0 = -.25.$$

By substituting the above values up to this—the fifth—order of differences,

$$y_x = 25 + (.4\times 3) + \frac{.4(.4-1)}{1\times 2}\times 1 + \frac{.4(.4-1)(.4-2)}{1\times 2\times 3}\times 0 + \frac{.4(.4-1)(.4-2)(.4-3)}{1\times 2\times 3\times 4}\times .5 + \frac{.4(.4-1)(.4-2)(.4-3)(.4-4)}{1\times 2\times 3\times 4\times 5}\times -.25.$$

$$= 25 + 1.2 + \frac{-.24}{2} + 0 + \frac{-.9984}{24}\times .5 + \frac{3.6}{120}\times -.25.$$

$$= 25 + 1.2 - .12 + 0 - .0208 - .0075$$

$$= 26.05.$$

The best estimate of the required annual premium payable at age 22 years is Rs. 26.05.

Interpolation for the value of y corresponding to any other x can also be made in a like manner.

Newton's formula uses differences running in a diagonal direction, and is suited for interpolation near the *beginning* of the table. It should be noted that it is used in a case in which the independent variable x advances by *equal increments.* In table 65, x advances by 5 years. When it is required to interpolate in a frequency distribution it is better to work with the cumulative numbers. For example, if from the frequency distribution of marks given in Table 22, page 170, it is desired to know the total number of candidates who obtained marks not exceeding 15, the Table, for purposes of calculating the differences and interpolating the number of students, would be written as follows (table 66), and interpolation carried out as above.

Table 66. *Cumulative Frequency of Marks in Economics.*

Number of marks x	No. of candidates y
Not more than 10	4
,, ,, ,, 20	12
,, ,, ,, 30	23
,, ,, ,, 40	38
,, ,, ,, 50	49
,, ,, ,, 60	56
,, ,, ,, 70	60

Lagrange's Formula.

When the recorded y's correspond to x's, and the x's advance by *unequal intervals,* the most convenient formula to use is that due to the famous French mathematician, Lagrange.

Representing the quantities as before by

$(x_0 y_0)$, (x_1,y_1), (x_2,y_2),..................(x_n,y_n),

Lagrange's formula runs thus:

$$y_x = y_0 \frac{(x-x_1)\ (x-x_2)\ldots\ldots\ldots\ldots(x-x_n)}{(x_0-x_1)\ (x_0-x_2)\ldots\ldots\ldots\ldots(x_0-x_n)}$$
$$+y_1 \frac{(x-x_0)\ (x-x_2)\ldots\ldots\ldots\ldots(x-x_n)}{(x_1-x_0)\ (x_1-x_2)\ldots\ldots\ldots\ldots(x_1-x_n)}.$$
$$+\ldots\ldots\ldots\ldots\ldots\ldots$$
$$+y_n \frac{(x-x_0)\ (x-x_1)\ldots\ldots\ldots\ldots(x-x_{n-1})}{(x_n-x_0)\ (x_n-x_1)\ldots\ldots\ldots\ldots(x_n-x_{n-1})}$$

where, y_x is the quantity to be interpolated, x is the given quantity in the x variable corresponding to which y_x is to be interpolated, $x_0, x_1, x_2, \ldots\ldots x_n$ are the given values of the variable x, and $y_0, y_1, \ldots\ldots y_n$ are the corresponding given values of the variable y.

The following table relates to income earned per month by a certain number of workers in a big manufacturing concern.

Table 67. *Monthly Income of Workers.*

Income not exceeding Rs.	x	Number of workers	y
15	x_0	36	y_0
25	x_1	40	y_1
30	x_2	45	y_2
35	x_3	48	y_3

It is required to estimate the number of workers getting not exceeding Rs. 26 per month.

Making use of the given data and taking $x=26$, we have,

$$y_x = 36 \frac{(26-25)\ (26-30)\ (26-35)}{(15-25)\ (15-30)\ (15-35)}$$
$$+40 \frac{(26-15)\ (26-30)\ (26-35)}{(25-15)\ (25-30)\ (25-35)}$$

$$+45\frac{(26-15)\ (26-25)\ (26-35)}{(30-15)\ (30-25)\ (30-35)}$$

$$+48\frac{(26-15)\ (26-25)\ (26-30)}{(35-15)\ (35-25)\ (35-30)}$$

$$=36\times\frac{1\times-4\times-9}{-10\times-15\times-20}+40\times\frac{11\times-4\times-9}{10\times-5\times-10}$$

$$+45\times\frac{11\times1\times-9}{15\times5\times-5}+48\times\frac{11\times1\times-4}{20\times10\times5}$$

$$=-\frac{54}{125}+\frac{792}{25}+\frac{297}{25}-\frac{264}{125}$$

$$=\frac{5127}{125}$$

$$=41$$

Therefore, income not exceeding Rs. 26 per month is being earned by 41 workers, as interpolated.

The above example relates to a case of frequency distribution where the magnitude of different class-intervals is not equal. If, instead of frequency distribution, individual items were given as, for instance, population of India during certain years, the years advancing not necessarily by equal intervals the method of attacking the problem would remain the same as that in the above example. In a like manner extrapolation can be made.

Forecasting.

While discussing graphic method of interpolation used in connection with the seasonal and cyclical series it was pointed out that when the cyclical character of a curve had been ascertained, it was easy to locate a missing figure, and it was also hinted there that such curves could prove useful for forecasting where economic events moved in a cycle. Periods of industrial boom or of agricultural depressions have been found to repeat themselves at an interval of 7—10 years. Therefore, when once it has been found, as a result of the study of sufficient and reliable data, that a certain phenomenon is characterised by cyclical tendency, its future course can be predicted, fairly accurately, upon the strength of past knowledge. This

prediction is nothing but forecasting. A practical businessman knows the facts of ups and downs in business from the experience; to him a periodicity or seasonal variation curve, or, forecast based on it, is of little interest. But to a statistician or to an economist, the knowledge of regular movement is of great assistance in predicting the course of many economic events. Even businessmen, particularly in the industrially advanced western countries, are making use of what is called Business Forecasting, which we will take up in the next Chapter.

Again, in Chapter XIII while dealing with logarithmic or ratio scale charts it was pointed out that such charts, once their fluctuating character had been studied, could be extended to predict a future figure relating to the phenomenon they illustrated. This is, again, done with a knowledge of the trend of the curve. In Figure 21, in that Chapter the dotted line shows the manner of extending the curve beyond the last date upto which evidence is available. This is a method of extrapolation, which leads to forecasting. From the extended portions of the curves in Figure 21 it can be predicted that the amount in 1943 would rise to Rs. 50.64 and Rs. 506.4 respectively in the two cases, A and B.

Conclusion.

After a discussion of the various methods of interpolation and extrapolation the practical utility of these methods must now be clear. In administration as well as in business, maintenance of minute records of each item from date to date is a matter of time, labour and money. It is impossible to take yearly census of population, for instance. But population figures are, year after year, necessary for estimating the income from existing taxes, for exploring the possibilities of new sources of revenue or for planning the production of food. How the methods, discussed above, are useful for estimating the population year after year, can be very easily seen. Similarly, a businessman's sales records may be incomplete or he may like to estimate the probable demand for his wares in the coming year. Methods such as those discussed above would provide him with the most likely estimate based on past experience. So, these methods are of immense practical use.

EXERCISES

(1) What are the assumptions that are made in interpolating missing figures in a series? How far are interpolated or extrapolated figures to be relied upon?

(2) Give a few examples of the use of Interpolation in Business Statistics.

(M. Com., Luck., 1942).

(3) Explain fully the process of interpolation by graphic method.

(4) What is interpolation? State Newton's formula for interpolation for equal intervals and the assumption under which it holds. Apply the formula to find the expectation of life at age 24 using the following data:—

Age (yrs.)	10	15	20	25	30	35
Expectation of life (yrs.) ..	35.4	32.2	29.1	26.0	23.1	20.4

(M.A., Raj., 1950).

(5) Estimate the value of imports in the year 1913 from the data given below. Use algebraic method:

Year		Value of Imports (Rupees)
1910	..	3,92,02,000
1911	..	2,65,10,000
1912	..	2,61,63,000
1913	..	..
1914	..	3,37,55,000
1915	..	3,29,87,000
1916	..	2,74,31,000

(M.A., Alld., 1950).

(6) From the following life-table, calculate the number living at ages 25, 35 and 47:—

Age (in years)	..	20	30	40	50
Number living	..	51	44	35	24

(M.A., Alld., 1952).

(7) The population of a town increases according to the compound interest law. In 1890 and 1940 it was 19,500 and 34,670 respectively. Sketch a graph to show the population at all times during the period, and use it to estimate—

(*a*) the population in 1926;
(*b*) the population in 1945; and
(*c*) the rate per thousand per annum of increase in population.

(M. Com., Alld., 1945).
(M.A., Raj., 1951).

(8) From the table given below, interpolate the number under 30 years. Use Lagrange's formula.

Ages (yrs.)		Proportion occupied per 10,000
10—15	..	193.5
15—20	..	880
20—25	..	933
25—35	..	1,636
35—45	..	1,201
45—55	..	830

(M.A., Alld., 1951).

(9) The following table gives the population of Gaya and Patna at the time of the last five censuses:

Years		Gaya	Patna
1891	..	80,383	165,192
1901	..	71,288	134,785
1911	..	49,921	136,153
1921	..	67,562	119,976
1931	..	88,005	159,690

Estimate the population of Gaya separately for the year 1941.

(M.A., Patna, 1943).

(10) (*a*) Give some uses of interpolation in Economic Statistics.

(*b*) The following table gives the quantities of a certain brand of tea sold at prices noted against each:

	Rs.	Rs.	Rs.	Rs.	Rs.
Price of tea per pound ..	1 4 0	1 8 0	1 12 0	2 0 0	2 4 0
Quantity sold (in thousand lbs.) ..	97.5	79.7	65.9	56.0	49.9

Estimate the probable sale when the price is Rs. 1-6-0 per pound.

(M.A., Patna, 1944).

(11) The age of mothers and the average number of children born per mother are given in the table below. Interpolate the average number of children born per mother aged 30—34.

Age of mother in years	Number of children born
15—19	0.7
20—24	2.1
25—29	3.5
30—34	?
35—39	5.7
40—44	5.8

(M. Com., Alld., 1946).

(12) Following are the marks obtained by 492 candidates in a certain examination:—

Not	more	than	40	marks	212	candidates
,,	,,	,,	45	,,	296	,,
,,	,,	,,	50	,,	368	,,
,,	,,	,,	55	,,	429	,,
,,	,,	,,	60	,,	460	,,
,,	,,	,,	65	,,	481	,,
,,	,,	,,	70	,,	490	,,
,,	,,	,,	75	,,	492	,,

Find out the number of candidates who secured more than 42 but not more than 45 marks.

(M.A., Cal., 1935).
(M. Com., Alld., 1947).

(13) Estimate the outflow of gold from India in 1937-38 from the data given below:—

Average value of net exports of gold coin and Bullion.

Year		Rs.
1933—34	..	89,56,32,418
1934—35	..	52,53,74,607
1935—36	..	37,35,58,955
1936—37	..	27,84,61,129
1938—39	..	23,26,02,068

(M. Com., Alld., 1948).

(14) Discuss the utility of interpolation and extrapolation to a businessman. What are the different methods known to you for interpolation?

(M. Com., Alld., 1950).

(15) From the following table, find the number of students who obtained less than 45 marks:—

Marks		No. of students
30—40	..	31
40—50	..	42
50—60	..	51
60—70	..	35
70—80	..	31

(M. Com., Alld., 1952).

(16) How will you calculate changes in intercensal population? Can you forecast the population in 1945 on the basis of census figures of population in 1921, 1931 and 1941 ?

(M. Com., Agra, 1945).

(17) Explain what you mean by 'Interpolation'. The following table shows the probable ages of bridegrooms for brides of various ages:—

Age of bride	Probable age of groom	Age of bride	Probable age of groom
15.5	25.0	25.5	27.0
16.5	25.2	26.5	27.5
17.5	25.4	27.5	28.0
18.5	25.5	28.5	29.0
19.5	25.5	29.5	30.0
20.5	25.5	30.5	32.0
21.5	25.8	31.5	33.0
22.5	26.0	32.5	33.5
23.4	26.0	33.5	34.0
24.5	26.8	34.5	34.5

Represent the figures graphically, and find out from the graph the probable age of the bridegroom for a bride aged 25.

(M.A., Agra, 1944).

(18) (*a*) What is interpolation and how is it useful?

(*b*) From the following table of Annual Premiums required to secure at death Rs. 1,000 with bonus under the Whole Life Assurance Policy of a certain Assurance Company, find by graphic method of interpolation the annual premiums for ages 34 and 36 nearer birthday respectively.

Age nearer Birthday	Annual Premium Rs. As.
25	27—14
30	31—13
35	37— 0
40	43—13
45	52—12

(B. Com., Andhra, 1944).

(19) From the following table of membership of a certain club, calculate the probable membership for 1943, assuming a linear law:—

Year ..	1931.	1932.	1933.	1934.	1935.	1936.
Membership ..	508.	524.	539.	556.	574.	590.
Year ..	1937.	1938.	1939.	1940.	1941.	1942.
Membership ..	604.	619.	638.	655.	671.	690.

Is your conclusion valid? If not, why not?

(B. Com., Andhra, 1943).

(20) State the assumptions underlying the finite differences formula employed in interpolation. Calculate the expectation of life at the age of 22 using the following data:—

Age		Expectation of life
10	..	35.45
15	..	32.20
20	..	29.06
25	..	26.03
30	..	26.11
35	..	20.41

(B. Com., Nagpur, 1943).

(21) Find an interpolated figure for population of 1896 from the following table:—

Year		Population
1881	..	25,974
1891	..	29,003
1901	..	32,528
1911	..	36,070

State the assumption underlying the formula employed.

(B. Com., Nagpur, 1942).

(22) The following table gives the different premiums at different ages in a Life Assurance Co.

Age		Premium (in Rs.)
25	..	23
30	..	26
35	..	30
40	..	35
45	..	42
50	..	51

Find the premium at age 28.

(B. Com., Nagpur, 1944).

(23) The annual sales of a concern are given below:—

Years		Sales of cloth in Lakhs of yards
1915	..	125
1920	..	163
1925	..	204
1930	..	238
1935	..	282

Assuming the conditions of the market to be the same, estimate the sales for the year 1940.

(M.A., Patna, 1941).

(24) Extrapolate the population of a town for 1946 from the following data about its population during the previous four censuses:—

Census year		Population in thousands
1911	..	473
1921	..	468
1931	..	454
1941	..	484

(M. Com., Raj., 1952).

(25) Interpolate the missing figure in the following table with the help of a suitable formula:

1911	..	1331
1912	..	1728
1913	..	2197
1914	..	?
1915	..	3375
1916	..	4096
1917	..	4913

(M.A., Delhi, 1953).

(26) What do you understand by 'extrapolation'? Under what circumstances is extrapolation used in statistics? Illustrate your answer.

(M. Com., Agra, 1945).

CHAPTER XIX

BUSINESS FORECASTING

Human conduct is based to a very large extent upon forecasts; any action that man takes in the present is taken in the hope that certain events will occur in the future and that they will take a certain shape. The basis for such expectation is the series of events which are known to have taken place in the past. Forecast of coming events on the basis of past experience is thus ingrained in human psychology.

Business Conduct.

What is true of human conduct is also true of business conduct. When a man enters business, he really enters a forecasting profession. His decision to take a risk is based on his estimate of the degree of risk involved and the reward for taking it, that is, on the expectation concerning the future. He has to frame the best possible estimate of the trend of interest rates, the future price of raw-material, the future demand for manufactured goods and so on. He has to have a peep into the future; he must forecast as a risk-taker. He may scorn the scientific business forecaster but cannot help being one himself. It may happen that he may not call to his aid the apparatus which the scientific expert holds so sacred and base his forecast on hunches, hearsay, customary belief, astrological predictions, impressions from trade associates, etc. It may also happen that if he is endowed with a shrewd natural instinct, or, if experience has sharpened his intellect, he may sense the future events fairly accurately. But, this does not mean that scientific business forecasting is useless or unnecessary. Forecasts based on hunches cannot always be true; nor are all businessmen gifted with shrewd instinct and sharp intellect. Therefore, for an ordinary businessman to conduct his business on proper lines, the application and study of the principles of business forecasting are extremely valuable.

What is Business Forecasting?

Business forecasting is not so much the estimation of certain figures of production, sales, profits, etc., as the statistical

analysis of past and current economic conditions with the object of determining probable future business conditions. It is concerned with drawing inferences from the known data, both internal and external, in a manner which will enable business policies to be determined to meet probable future conditions to the best advantage. The whole problem, therefore, has two aspects:

(1) **Analysis of past business conditions.**—Historical analysis shows the course that business has followed and there is great advantage in knowing the nature of long-period trend, the extent of seasonal, cyclical and irregular fluctuations, the correlation between time series, the extent of lag, etc.

(2) **Analysis of current economic data.**—This analysis is done in relation to a probable future tendency. The intention is to determine the factors that may alter sequental changes observed by historical analysis. Examples of such factors are new inventions and discoveries, changes in public tastes and fashions, politico-economic changes, devaluation of country's currency, raising a tariff wall, Government's economic policies, war, armistice, etc. All such factors have an important bearing on the coming future. Their influences have, therefore, to be borne in mind when drawing conclusions or attempting to forecast the probable course of business on the basis of historical analysis.

Both the above aspects are equally important in business forecasting. Business activity is subject to successive waves of prosperity and depression. Historical analysis would reveal these waves. But, cycles of prosperity and depression do not recur at regular intervals; they vary both in their intensity and duration inasmuch as they are the resultant of a complex group of causes. Under the circumstances, forecasting is made possible only by the knowledge of the extent to which similar conditions are found to recur and of the signs indicating the time when the recurrence is likely to take place. From detailed studies of trade cycles in several countries, it has been discovered that although changes in business conditions are neither fortuitous nor periodic or regular, yet, in general, business passes through well-defined major and minor changes. Accordingly, it it possible to study the order of changes in different businesses and to measure their present related movements. These movements serve as good indicators for deter-

mining when economic conditions are basically sound or unhealthy and when business prosperity is likely to undergo a downward swing. For example, the Harvard Committee on Economic Research in USA found after an elaborate study of the data for the period 1903 to 1914 that movements in Speculation preceded those in Business while the latter preceded those in Money. These movements occurred with such marked regularity of sequence that they afforded a logical basis for scientific business forecasting: occurrence of a particular movement in Speculation, for instance, meant that a change in Business was impending and that, thereafter, the change would be passed on to Money. The Harvard Committee, however, did not end its job with the study of the past; it studied current economic situation too, to find out whether there were any factors which, in future, would affect the sequence observed in the past.

Statistical Data and Methods used in Forecasting.

From the above it will be seen that the basic data required for business forecasting are statistical series indicating the movements of production, commercial, financial and other activities in the past and the present. One must know the volume and value of goods produced or exchanged, volume and value of foreign trade, amounts represented by banking transactions, volume of money in circulation, prices per unit of different commodities, etc.

The original data are converted into index numbers for convenience of comparison. The index series are plotted on a graph, and analysed so as to eliminate their trend, seasonal and cyclical fluctuations. Cycle percentages are then calculated by dividing the cyclical variations by the standard deviation of the series, as explained in Chapter XV on Correlation while dealing with "Coefficient for Cyclical Fluctuations". Time lag between two related series is then discovered, by either superimposing cycle percentages curve of one series over the other, or, by calculating with different lags the coefficients of correlation between cycle percentages. That time lag which gives the maximum value for the coefficient of correlation between cycle percentages of two series is the best fitting lag. The

coefficients of correlation of cycle percentages of a great number of time series, in fact, provide a basis for selecting those which lag behind or precede other series. It was in this manner that Persons constructed from the annual data of a large number of time series both a business barometer and a forecaster. The same method, elaborated, refined and applied to data for 1903-1914 laid the foundation for the business barometric and forecasting lines of the Index of General Business Conditions issued by the Harvard Committee on Economic Research in U.S.A.

Having discovered the best fitting lag between pairs of time series, they may be plotted for a sufficiently long period of time, allowing for the lag. From the resulting correlation graph, it will be possible to study the sequence in the movements of the different business and economic series, and on the basis of this knowledge to forecast the probable trend of future business conditions. This knowledge will, however, have to be used along with the knowledge of current economic events, so that, while forecasting, the inferences flowing from the past are modified in the light of the current economic conditions.

Business Index or Business Barometer.

One of the modern devices employed in scientific business forecasting is that of the Business Index, also called Business Barometer or Economic Forecaster. An index of this type may refer to general conditions of trade or industry or to a particular trade or industry or even to an individual business. Graphical presentations of index numbers of volume and value of goods produced, imported and exported, of financial statistics relating to bank clearings and money rates, of prices of stocks and shares, of wholesale prices of commodities, of wages, of electrical power—and these and similar other indices in the field of production, commerce, finance, prices, etc.—give a readier view of their long-period trend and seasonal variations which can be eliminated through the statistical methods. Having done so, attention is separately concentrated on trend, cyclical variations and seasonal oscillations, to examine the underlying tendencies and related movements of the different series. The way is thus prepared for logical inference.

Index numbers relating to different activities in the field of production, trade, finance, etc., may also be combined into a general Index of Business Activity. They will have to be weighted suitably. This general index would refer to general conditions of trade and industry. But the behaviour of individual industries or trades might show a different trend from that of the composite Business Activity Index. Also, general boom or depression may be reflected in a majority of separate industries and trades, yet some industries and trades might show quite contrary tendencies. It follows, therefore, that study of general business conditions as revealed by the general Business Index should be supplemented by special studies of individual businesses based on separate indices.

Several trade and business organisations issue reports and charts relating to general business conditions or specific aspects of business. Many banks and business institutions publish statistical reports indicating various aspects of business outlook. But not all businesses follow the general trend and consequently barometers dealing with important industries and trade are prepared separately and published. These index numbers or barometers afford guidance when considering sales campaign or making bulk purchases. If the barometers indicate that prices have a fair chance of rising in the coming future, the businessman may plan a sales campaign; if, on the other hand, the indications are that they would fall, he may plan for bulk purchases. Similarly, the trends indicated by barometers may guide a businessman in regard to the advisability of embarking on heavy capital investment.

Theories of Business Forecasting.

Several theories of business forecasting are now known and as researches take place more theories may come to light. We take up below the discussion of four such theories.

Time-Lag or Sequence Theory.

This theory is based on the fact that, ignoring minor fluctuations, cyclical movements are not simultaneous but successive. For instance, if there is currency inflation, it may be followed by a rise, in order, in exchange rate, wholesale prices,

retail prices, cost of living, money wages. Contrary-wise, currency deflation may be followed, in order, by fall in exchange rate, wholesale prices, retail prices, cost of living and money wages. Rise or fall in wholesale prices is followed by a corresponding change in production and commercial activity. The sequence mentioned above is generally maintained, although time-lag between different movements varies with the rate of appreciation or the rate of depreciation of currency.

It was on the basis of the study of time-lag and sequence between Speculation, Business and Money that the Harvard Committee founded its forecasting service. A certain sequence may be found to exist as between different industries or even as between different countries. If the jute mills in United Kingdom were dependent on India and Pakistan for the supply of raw-jute, it is natural that the output of jute manufacturers in U.K. will be related in some way to the production of raw jute in India and Pakistan.

Action and Re-action Theory.

The basis of this theory is that for every action there is an equal and opposite reaction. If the wholesale prices show an upward swing, the supply of the goods concerned may be augmented in due course, and this would bring a fall in the prices.

Cross-Cut Analysis Theory.

This theory denies that history repeats itself in economic life. According to this theory each situation is taken independently and influence of each of the factors affecting the situation studied.

Specific Historical Analogy Theory.

According to this theory the business forecasting agency selects, as a basis for forecasting, a past period resembling the present as far as possible. For example, given the weekly rainfall data for a certain region over a very long period of time, it is possible to pick out of the long series a particular year in which the extent and distribution of rainfall was as good or as bad as in the current year. Then, a knowledge of

the size of outturn of crops in that particular past year may afford a reliable guidance for forecasting the crop prospects in the current year.

Underlying Assumptions.

In all the above theories, except the Cross-Cut Analysis Theory, the fundamental assumption is the belief in an orderly development of things, in a slow but steady change, in a recurrence of events. The recurrence of events, however, is not taken to occur mechanically; new factors are also accounted for and their influence studied. The Cross-Cut Analysis Theory does not draw inspiration from the past; it studies each phenomenon on its own merits.

For forecasting, unless new factors likely to cause a change begin to operate, it is assumed that the line of trend determined by the data for recent years may be projected forwards, at any rate, for periods in the immediate future. This projection may be done either freehand in keeping with the past trend, or with the help of mathematical formulæ.

Utility of Business Forecasting.

The utility of business forecasting to the businessman in particular and society in general cannot be over-emphasised. Business cycles are generally disruptive in their effects upon the general economic organisation. Sudden and rapid price changes render uncertain the pecuniary calculations of businessman. Consequently risks of industry are greatly increased, business enterprise is deterred, growth of capital is discouraged, speculation is induced and international trade and financial relations are disturbed. When the evil effects of business cycles spread from one country to another, the entire economic machine of the world may be thrown out of gear, as it happened in the thirties of this century. If, therefore, a slump is forestalled, economic disaster can be avoided, or, at least, successfully faced. If fore-warned, the businessman will keep himself ready to weather the storm when it comes. Similarly, when he expects a boom he will take steps in advance to prepare himself for reaping a good harvest. Planning ahead of a boom or depression on the part of businessman will lessen

his risks, it will minimise economic disorder and will not disturb too much the economic stability. This is in the larger interests of the society.

Business forecasting is also of great utility to the merchant, investor, railways, banks, insurance companies, etc. A merchant is interested in the likely course that prices would take in the near future to plan his purchase. The investor is likewise interested in knowing the probable course, the prices of stocks and shares might take. Similar are the interests of other business institutions.

Business forecasting is not only useful to the businessman; it is of equally great value to the administrator. The administrator warned in advance about the coming trade cycle can plan to control it, or, at any rate, to lessen its irksome effects. The evil consequences of trade cycles on various economic phenomena and, therefore, on the political and economic stability of the country at large can be kept by the administrator at the lowest ebb, only if he is told of what is coming.

Thus, business forecasting helps in profit making or avoiding loss by administering a warning about the major changes in trade or business that are about to occur in the near future. It helps in reducing the intensity of trade cycles by enabling the businessman to lessen the range of price fluctuations. It enables the administrator to control the trade cycles, through devices like regulation of currency and credit.

The usefulness of scientific business forecasting has prompted many institutions in different countries of the world to institute business forecasting services. Among these may be mentioned the Harvard Economic Society, the Brookmire Economic Service, Babson's Statistical Organisation, the Standard Statistics Corporation and the Annalist's Service in the U.S.A. and the London and Cambridge Economic Service in Great Britain. In India, index numbers of industrial production, quantity and value of imports and exports, wholesale prices, security prices, cost of living, factory earnings, money supply, etc., are now available. But no co-ordinated attempt has yet been made to utilise these data for business forecasting. One reason for this probably is that some of these index numbers have been started newly and long series over a period of

time in respect of them are not available. Another reason may be the inadequacy of the index numbers, inasmuch as some of them are not all pervading; for instance, the index number of industrial production does not cover all industries in the country. It also needs an agency interested in business forecasting to take up this work.

Limitations.

It is necessary to bear in mind the limitations of business forecasting and the caution that is necessary in using its results. Forecasting by itself cannot provide sure road to success; yet it is one element, and an important one, in carrying business in general to a successful end. Forecasting is far from being a precise science. It is still developing. Economic conditions, no doubt, tend to repeat themselves, but other factors appear on the scene and prevent exact repetition so that a clear view of the future is not always possible. The results of forecasting can, at best, be taken as indications of what is likely to come and nothing more.

EXERCISES

(1) What is Business Forecasting? What different theories of it are known to you? Explain clearly any two of them.

(M. Com., Alld., 1945).

(2) What is meant by Business Forecasting? Explain the major classes of methods used in forecasting.

(M. Com., Lucknow, 1944).

(3) What is meant by Business Forecasting? How far do Economic Barometers help the forecasting of economic events?

(M. Com., Alld., 1948).

(4) Discuss the important theories of Business Forecasting. How does analysis of time series help in forecasting of economic events?

(M. Com., Alld., 1952).

(5) What is meant by a Business Activity Index? How will you weigh the various series of which such index will be composed?

(M. Com., Lucknow, 1943).

(6) In what field of business facts must be studied to judge the position of a business in regard to Business Cycles? Why is a careful analysis of the business cycle important in the field of marketing.

(M. Com., Lucknow, 1943).

(7) Examine and illustrate the importance of Business Forecasting in the modern world. What are its limitations?

(8) Methods of extrapolation in Statistics are important in themselves; they are valuable also because they make possible the forecasting economic events. Discuss.

(9) Explain the place of time-series analysis in the sphere of business forecasting.

CHAPTER XX

PROBABILITY AND SAMPLING

PROBABILITY

In Chapter XIII the forms which frequency distributions take were described. Reference was also made there to the symmetrical bell-shaped curve called the Normal Curve of Error. This curve is of the greatest importance in the Science of Statistics inasmuch as many of the measurements used to describe distributions of observations (*e.g.*, the mean, the standard deviation, the co-efficient of variation) are distributed in accordance with this curve. In sample studies, the results obtained from the sample are generalised for the population. The Normal Curve of Error helps a great deal in determining the reliability of such generalisation. The basis of the Normal Curve of Error, however, is the Theory of Probability which we may deal with first.

Elementary Principles of Probability.

Reference was made to the principle of Probability in Chapter XVII on Association of Attributes, while explaining how the fact of association between two attributes can be established by comparing observation with expectation. That discussion will be expanded into somewhat greater detail here.

If an event can occur in n ways of which a are considered as successful and b as unsuccessful, the probability p of a successful outcome may be written as under:—

$$p=\frac{a}{n}$$

The probability q of an unsuccessful outcome is

$$q=\frac{b}{n}$$

Since the sum of the successful and unsuccessful outcome is equal to the total number of events, we have

$$a+b=n$$

$$\text{or, } \frac{a}{n} + \frac{b}{n} = 1$$

so that, $p+q=1$, or *certainty*

Simple Probabilities.

If a coin is tossed, it should fall with either the head or the tail upwards. If the turning up of a head is taken as a successful outcome, the probability p of a success is

$$p=\tfrac{1}{2}$$

and the probability q of a failure is

$$q=\tfrac{1}{2}$$

Similarly, if a card is drawn from a pack of 52, the chance of drawing the jack of spades is $\frac{1}{52}$ and the chance of failing in that endeavour is $\frac{51}{52}$

Likewise, if we throw a die, the probability p of getting a six spot is

$$p=\frac{1}{6}$$

and the probability q of failing to get a six spot is

$$q=\frac{5}{6}$$

Addition of Probabilities.

If a single draw is made from a pack of 52 cards what is the chance of securing either an ace of spades or a three of spades? In such a case the probability of a success is equal to the sum of the separate probabilities.

The probability of an ace of spades is $\frac{1}{52}$. Also, the probability of a three of spades is $\frac{1}{52}$. Therefore, the probabiltiy p of securing *either* an ace of spades *or* a three of spades is

$$p=\frac{1}{52}+\frac{1}{52}=\frac{1}{26}$$

Similarly, the probability of drawing *either* a club *or* a heart from a pack of 52 playing cards is

$$p=\frac{13}{52}+\frac{13}{52}=\frac{1}{2}$$

Multiplication of Probabilities.

The probability of a compound event is the product of the probabilities of the separate events. A compound event is one where, say, two events, independent of each other, will both occur. In two successive throws of a die, the chance of securing a two spot followed by a three spot is

$$p=\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$$

Addition and Multiplication of Probabilities.

In estimating the probability of a given outcome it is often necessary both to add and multiply probabilities. Suppose we are to determine the chance of securing the total 5 from two dice thrown simultaneously, the dice being labelled a and b to distinguish them. The total of 5 can be secured from any one of the four following combinations.

die a	1	2	3	4
die b	4	3	2	1
total	5	5	5	5

The chance of securng an ace or one spot with die a is $\frac{1}{6}$ and of securing a four spot with die b is $\frac{1}{6}$. The chance of the

two in combination is $\frac{1}{36}$ Similarly, the probability of securing a two spot with die *a* is $\frac{1}{6}$ and that of securing a three spot with die *b* is $\frac{1}{6}$ so that the chance of the two events in combination is $\frac{1}{36}$. Same is the case with each of the other two combinations. But any one of these four results will give a total of 5. Hence the probability *p* of getting a total 5 from two dice thrown simultaneously is

$$p=\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}=\frac{4}{36}=\frac{1}{9}.$$

The question that we have answered above is as to what the probability is of securing a total of *exactly* 5 in the toss of two dice. If, on the other hand, the question is as to what the probability is of securing a total of *at least* 5 in the toss of two dice, a total of 5 or more will be considered a successful outcome. As in the above example, we will then work out the probability of securing each of the successful results, that is, all results above and including the throw of 5. We will find that with two dice the probability of throwing 12 is $\frac{1}{36}$, that of throwing 11 is $\frac{2}{36}$, that of throwing 10 is $\frac{3}{36}$ and so on. If we calculate all such probabilities for a total of 5 or more, and add them up, we have

$$p=\frac{1}{36}+\frac{2}{36}+\frac{3}{36}+\frac{4}{36}+\frac{5}{36}+\frac{6}{36}+\frac{5}{36}+\frac{4}{36}$$
$$=\frac{30}{36}$$

Thus, the chance of throwing *at least* 5 in the toss of two dice comes to be $\frac{30}{36}$ or $\frac{5}{6}$.

The Binomial Expansion.

The above facts can be expressed in a generalised form aalso. If two rupee coins a and b are tossed simultaneously, there are four possible outcomes:

a b	a b	a b	a b
H H	T H	H T	T T

where H stands for the head and T for the tail. The first outcome is that both the rupee coins a and b may fall with heads upwards; the second is that a may fall with head up and b with tail up; the third is that a may fall with tail up and b with head up; and, the fourth is that both a and b may fall with tails up. In other words, the chance of securing 2 heads is 1 out of 4 outcomes, that is, it is $\frac{1}{4}$; that of securing 1 head is $\frac{2}{4}$ or $\frac{1}{2}$; and, that of securing no heads is $\frac{1}{4}$.

Let us suppose that the falling of head upwards is a successful event and that of tail upwards is an unsuccessful event. Further suppose that p stands for the probability of success of an event and q for the probability of its failure. Then, the above four outcomes can be written as under:

$$\begin{array}{cccc} HH & HT & TH & TT \\ pp & \underbrace{pq \quad qp} & & qq \\ \text{or, } p^2 & 2pq & & q^2 \end{array}$$

Now, $(p^2+2pq+q^2)$ is equal to $(p+q)^2$. Therefore, if there are two independent events, their combined probability is given by the expansion of the expression

$$(p+q)^2$$

If $p=q=\frac{1}{2}$, then the probabilities of the various outcomes (that is, 2 heads, 1 head and no heads respectively) are ,

$$(\tfrac{1}{2}+\tfrac{1}{2})^2=\tfrac{1}{4}+\tfrac{1}{2}+\tfrac{1}{4}.$$

This is exactly what we got above.

Likewise, if three rupee coins a, b and c are tossed simultaneously, there are 8 possible outcomes, and they will give the expansion of the expression

$$(p+q)^3$$

Thus,

$$\begin{array}{ll} HHH = p^3 & = p^3 \\ \left.\begin{array}{l} HHT = p^2q \\ HTH = p^2q \\ THH = p^2q \end{array}\right\} & = 3p^2q \\ \left.\begin{array}{l} THT = pq^2 \\ TTH = pq^2 \\ HTT = pq^2 \end{array}\right\} & = 3pq^2 \\ TTT = q^3 & = q^3 \end{array}$$

It will be seen from the above that the chance of securing three heads is 1 out of 8 outcomes, that is, $\frac{1}{8}$; that of securing two heads is $\frac{3}{8}$; that of securing one head is $\frac{3}{8}$; and, that of securing no heads is $\frac{1}{8}$. Now, if $p=q=\frac{1}{2}$, we have

$$p+q)^3 = p^3+3p^2q+3pq^2+q^3$$
$$=\tfrac{1}{8}+\tfrac{3}{8}+\tfrac{3}{8}+\tfrac{1}{8}$$

It should now be clear that if we wish to know the probable frequencies of the various outcomes in a given number of trials, we may use the expression

$$N\ (p+q)^n,$$

where N stands for the number of trials and n for the number of independent events. Thus, if there are 100 trials and there are two independent events, the probable frequencies are

$$100(p+q)^2=100(p^2+2pq+q^2)$$

If p is equal to $\frac{1}{2}$ so that q is also equal to $\frac{1}{2}$, then, we have,

$$100(\tfrac{1}{4})+100(\tfrac{1}{2})+100((\tfrac{1}{4})=25+50+25$$

which gives the probable frequencies of two successes, one success and no successes. In order words, if there are 100 trials and there are two independent events, the probability of 2 successes is 25, that of one success 50, and that of no successes 25.

Similarly, if there are three independent events, probable frequencies in N trials are given by the binomial expansion of

$$N(p+q)^3$$

By this expansion we will get, in order, the probable frequencies of three successes, two successes, one success and no successes.

It follows from the above that if we know in advance the probabilities attached to similar independent events, we are in a position to determine the probable frequencies of any given number of successes or failures. This is so whether p is equal to q or is not equal to q. What is necessary is that p and q should remain constant. This is a significant point which should be borne in mind.

Comparison of Actual and Theoretical Frequencies.

Six dice were thrown 64 times. Each 4, 5, or 6 spot appearing was considered to be a success, and each 1, 2 or 3 spot a failure. The results were as recorded in Table 68 col. (2).

Let us now compare with these results those which might be expected from the given conditions. Since six dice were thrown each time, we are dealing with six independent events. There are 4 trials. Since either a 4, 5 or 6 spot is considered a success, $p=q=\frac{1}{2}$. Now, our binomial expression is

$$N\ (p+q)^n.$$

For the terms in the binomial expansion, we have,

$$(p+q)^n=p^n+np^{n-1}q+\frac{n(n-1)}{1.2}p^{n-2}q^2+\frac{n(n-1)\ (n-2)}{1.2.3.}p^{n-3}q^3 +\ldots\ldots+q^n.$$

In the case of our example, the binomial expansion is given by

$$N(p+q)^6$$

Now,

$$N(p+q)^6=64\ (p^6+6p^5q+15p^4q^2+20p^3q^3+15p^2q^4+5pq^5+q^6)$$

$$=64\left(\frac{1}{64}+\frac{6}{64}+\frac{15}{64}+\frac{20}{64}+\frac{15}{64}+\frac{6}{64}+\frac{1}{64}\right)$$

Completing the indicated multiplication, we get the theoretical frequencies of the various possible successes in 64 throws of 6 dice. These are shown in col. (3) of Table 68.

Table 68. *Comparison of Actual and Theoretical Frequencies in throwing six dice.*

(1)	(2)	(3)
No. of Successes	Actual or Observed Frequencies	Theoretical or Expected Frequencies
0	0	1
1	5	6
2	13	15
3	22	20
4	17	15
5	7	6
6	0	1
	64	64

Mean & Standard Deviation of Theoretical Frequencies.

The arithmetic mean of the observed frequencies in Table 68 is 3.125 and their standard deviation is 1.1. The arithmetic mean and the standard deviation of the theoretical frequencies can be determined in a very simple manner, when we are aware of the probabilities involved. The formula used is

$$M=np$$

where M stands for the arithmetic mean, n for the number of independent events and p for the probability of success.

In our example in Table 68,

$$M=6\times\tfrac{1}{2}=3$$

Similarly, the standard deviation of the frequency distribution when the probabilities are known can be computed by the formula

$$\sigma=\sqrt{npq}$$

In the case of our example,

$$\sigma=\sqrt{6\times\tfrac{1}{2}\times\tfrac{1}{2}}$$
$$=\sqrt{1.5}$$
$$=1.22$$

The Normal Curve of Error.

The distributions of observed and theoretical frequencies in Table 68 are shown in Figure 34. How close the two frequency curves are to each other can be seen from the Figure.

The curve of theoretical frequencies in Figure 34 is a perfectly symmetrical frequency polygon. It is a six-sided polygon since the number of events is six. With 30 events, we will get a thirty-sided figure. With n infinitely large, a perfectly smooth curve would result. This curve of theoretical frequencies is the Normal Curve of Error, or a curve showing normal distribution. The normal curve is also known as the Gaussian Curve.

In the above example $p=q$. Therefore, skewness is zero. **Even if p and q were not equal in the above example, the binomial expansion would have remained unchanged, but the frequency curve would have been a skew one.**

Theoretical frequencies can be determined with the help of certain equations also, without undergoing the laborious process based on the binomial expansion and illustrated in Table 68. The normal curve has two equations. They depend on the area taken by the curve. One equation is

$$y=y_0e^{-\frac{x^2}{2\sigma^2}}$$

where y_0 is the maximum ordinate and is a constant; e is a constant having a value of 2.71828; σ is the standard deviation; and, x is a given value of the dependent variable expressed as a deviation from the mean. The maximum ordinate may be derived from

$$y_0=\frac{N}{\sigma\sqrt{2\pi}}$$

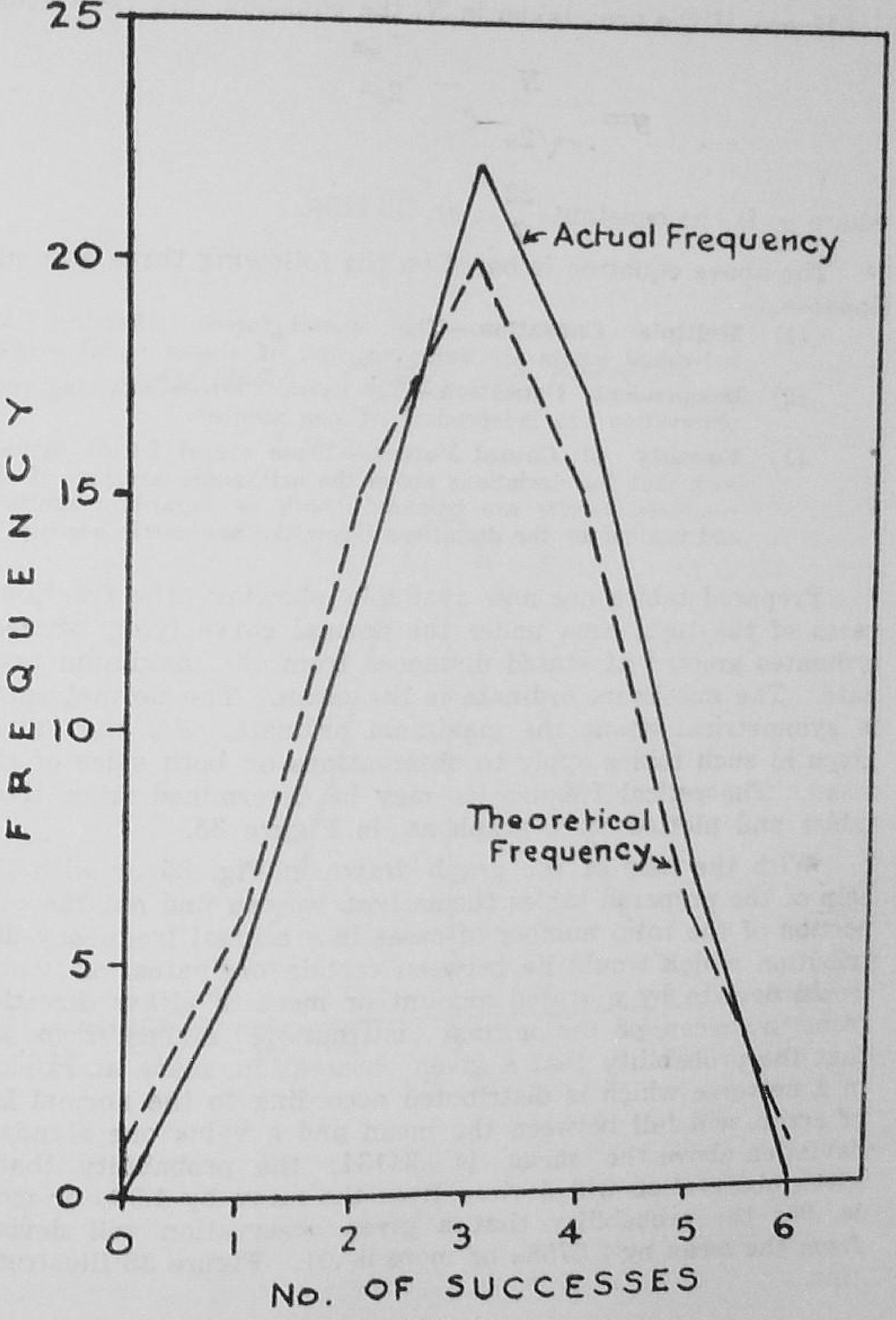

25
20
15
10
5
0
FREQUENCY
Actual Frequency
Theoretical Frequency
0
1
2
3
4
5
6
No. OF SUCCESSES

Fig. 34

Hence, if the area taken is N, the equation may be written

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

where π is the constant $\frac{22}{7}$, or, 3.14159.

The above equation is based on the following three assumptions:—

(1) **Multiple Causation.**—The causal forces affecting the individual events are numerous, and of almost equal weight.

(2) **Independent Causation.**—The causal forces affecting each observation are independent of one another.

(3) **Equality of Causal Forces.**—These causal forces operate such that the deviations above the arithmetic average of the combined results are balanced both as regards magnitude and number by the deviations below the arithmetic average.

Prepared tables are now available showing the fractional parts of the total area under the normal curve lying between ordinates erected at stated distances from the maximum ordinate. The maximum ordinate is the mean. The normal curve is symmetrical about the maximum ordinate. So, the values given in such tables apply to observations on both sides of the mean. Theoretical frequencies may be determined from these tables and plotted on a graph as, in Figure 35.

With the help of the graph drawn in Fig. 35 or with the help of the prepared tables themselves, we can find out the proportion of the total number of cases in a normal frequency distribution which would lie between certain ordinates, or, which would deviate by a stated amount or more in *either* direction from the mean of the normal distribution. Suffice it to say that the probability that a given observation, made at random in a universe which is distributed according to the normal law of error, will fall between the mean and a value one standard deviation above the mean is .34134; the probability that a given observation will deviate from the mean by 1.96σ or more is .05; the probability that a given observation will deviate from the mean by 2.5758σ or more is .01. Figure 35 illustrates this.

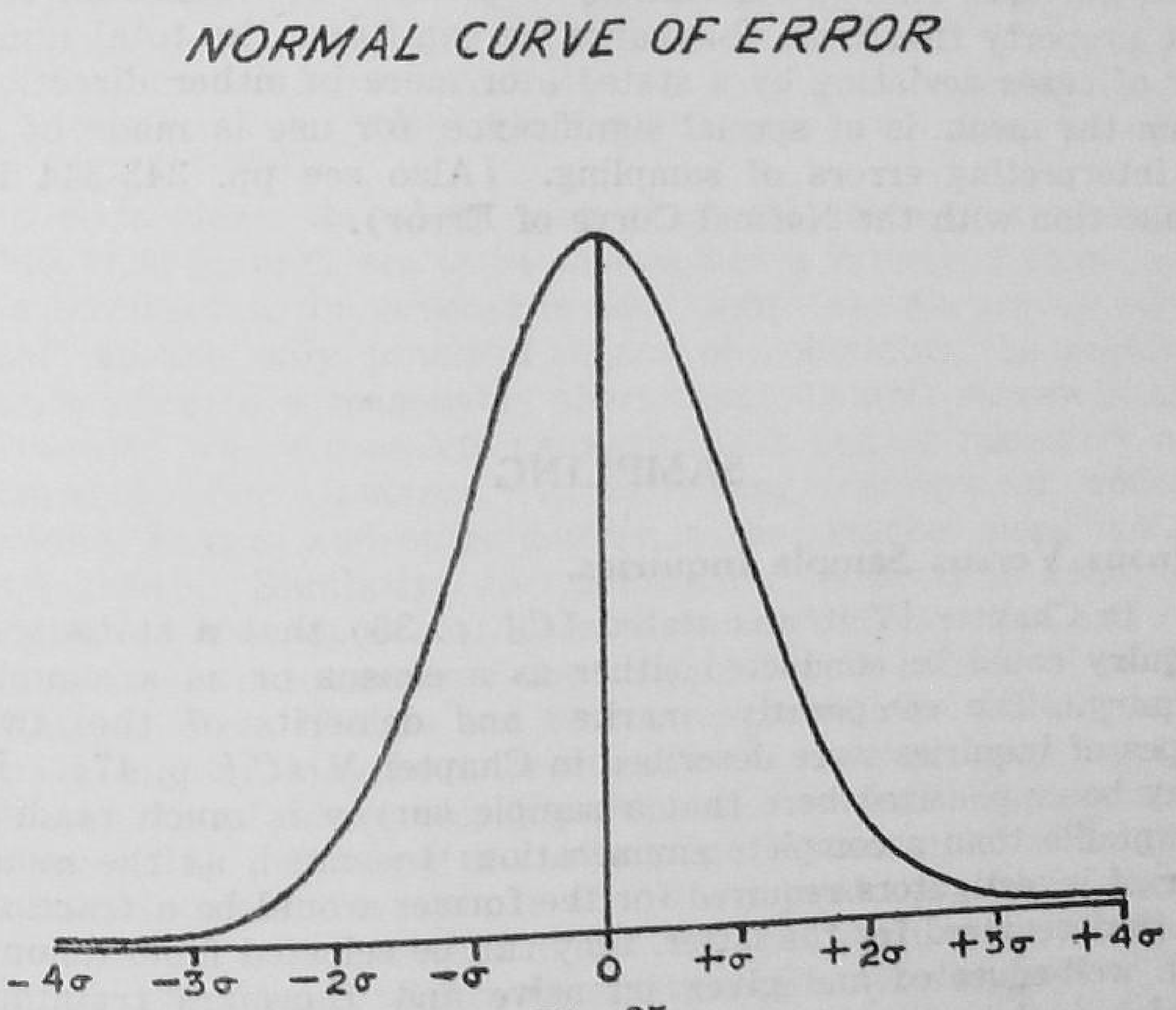


Fig. 35

From Fig. 35 it will also be seen that at the point 0 lies the mean, and that at that point the length of the ordinate is the highest. The mode, the median and the mean all coincide at this point. The limit of the curve, it will be seen, is infinity. For practical purposes, however, it is $a \pm 3\sigma$ which includes nearabout 97% of the cases. Almost 95% of the cases are covered by $a \pm 2\sigma$ and about 68% of the cases by $a \pm \sigma$.

The normal curve of error is one of the most significant curves in Statistics. It serves as a reference to measure skewness, kurtosis, standard deviation, coefficient of variation, etc. The property that from this curve we can know the total number of cases deviating by a stated σ or more in either direction from the mean, is of special significance, for use is made of it in interpreting errors of sampling. (Also see pp. 343-344 in connection with the Normal Curve of Error).

SAMPLING

Census Versus Sample Inquiries.

In Chapter IV it was stated (*C.f.* p. 36) that a statistical inquiry could be conducted either as a census or as a sample inquiry. The comparative merits and demerits of the two types of inquiries were described in Chapter V (*C.f.* p. 47). It may be emphasized here that a sample survey is much readily adaptable than a complete enumeration: Inasmuch as the number of investigators required for the former would be a fraction of that required for the latter, they can be selected from among the well-educated and given intensive and thorough training, which, in turn, would ensure a reliable and efficient conduct of the inquiry It may also be reiterated that a sample survey would require far less time and money to conduct than a census. Besides these advantages which a sample survey enjoys, it is significant to note that it is acclaimed as much more scientific than complete enumeration. During the course of his Presidential Address on "The Sub-Commission on Statistical Sampling" at the session on Sampling at the meeting of

the International Statistical Institute, Berne, in September, 1949, Prof. R. A. Fisher said:*

> "I have made four claims for the sampling procedure. About the first three, adaptability, speed and economy, I need say nothing further. Too many examples are already available to show how much the new method has to give in these ways. But, why do I say that it is more scientific than the only procedure with which it may sometimes be in competition, the complete enumeration? The answer, in my view, lies in the primary process of designing and planning an enquiry by sampling. Rooted as it is in the mathematical theory of the errors of random sampling, the idea of precision is from the first in the forefront. The director of the survey plans from the first for a predetermined and known level of precision; it is a consideration of which he never loses sight; and the precision actually attained subject to well understood precautions is manifest from the results of the enquiry."

It is clear, therefore, that where the scope of a survey is wide, that is, data are to be collected on a variety of items, and the territory to be covered is also large, sample survey offers itself as the only practical means of obtaining the required information in a reasonably short time. Sample survey is also advocated where complete enumeration is neither necessary nor desirable. For instance, in marketing surveys all sellers, brokers, buyers and other dealers in the market need not be interviewed. Similarly, no investigator desiring to know the heights of the population of India can afford the time and expense required to measure the height of each individual in the country. Even if time and resources permit, the conduct of a census inquiry may be impracticable in certain cases, as, for instance, those where only a portion of the whole universe may be accessible. The contents of a mine, for example, cannot be examined unless the mine is completely used up. Also, where the universe is heterogeneous and does not conform to the normal law of error, it will be necessary to survey every unit of it, that is, to conduct a census, in order to be able to draw reliable inferences.

The above is, however, not to say that sample survey is a complete substitute for census inquiry. Where one in interested in minute details or in the characteristics of individual con-

* Cited in the National Sample Survey, General Report So. 1 on the First Round, October 1950—March, 1951, p. 4.

stituents of a population, sample survey is ruled out. No sample survey of population would, for instance, yield information on the number of people under different age groups living in each village of India; only a complete enumeration would disclose this information.

A complete census can, however, be taken only at periodic intervals on grounds of, say, time, resources and problems of organisation. Sample surveys, conducted during the intercensal years, would greatly help in keeping the data on at least the main items of information uptodate. Sample surveys can also be utilised to check up the precision of the data collected through the census, as was done in India in the case of the 1951 population Census.

Objects of Sampling.

The fundamental object of sampling is to get as accurately as possible a picture of the whole universe by examining only a part of it. An attempt is thus made through sampling to give the maximum information about the parent universe with the minimum effort.

Another object of sampling is to determine the reliability of the estimates once obtained. This can be done by drawing successive samples from the same parent universe and comparing the results obtained from different samples, or, by making a sample check of a complete census conducted earlier.

Essentials of Sampling.

A sample should be so selected as to be representative of the entire universe from which it is drawn. In order to be representative of the universe as a whole, it should contain characteristics in proportion to their presence in the universe.

Secondly, the number of samples should be such that variation between them would lie between explainable limits.

Lastly, the size of a sample should be such as to enable a proper estimate of the sampling error.

Methods of Selecting a Sample.

In Chapter V two methods of selecting a sample were mentioned, *viz.*,

1. Purposive Sampling, and
2. Random Sampling.

The application and the respective merits and demerits of these methods were also described in that Chapter (*Cf. pp.* 46-47). It would be well to remember here that while both the methods aim at selecting a representative sample from the given universe, the selection is not quite objectively done in the former. In other words, while there is ample chance for the personal discretion of the organiser of a purposive sample to influence his selection, no such chance exists in a random sample. A random sample, by definition, is one where each unit in the universe has the same chance of being chosen; that is, it is entirely objective. Personal bias in a purposive sample influences the whole result in the same direction; that is, its effect is cumulative. In a random sample, the influences cancel each other; they are non-cumulative or compensating. It is also significant to remember that while the results obtained from a random sample can be assessed in terms of probability, that is, while we can measure the errors of estimation or the significance of results obtained from a random sample, the same is not possible in the case of purposive sampling. This, in fact, accounts for the preference shown by statisticians for random sampling.

Sometimes a combination of the above two methods may also be tried. In that case, the given universe may be divided purposively into homogeneous groups, and from each group the sample may be selected randomly. This method is referred to as Mixed Sampling method.

Bias.

Selection of a sample by any of the above-mentioned methods may be affected by what is known as bias, that is, perturbations which affect the choice of the sample and make it something different from what the investigator intends it to be. Bias may arise in several ways. It has already been said above that a purposive sample may be biased owing to the personal discretion of the investigator influencing the selection of the sample. Thus, in a survey of the yield of crops all district officers may tend to select fields whose yield is much more than

the average, in order to glorify their district. Such bias, however, does not enter a random sample. In a random sample, there are several other sources of bias.

Firstly, there is the bias in the *selection* of a random sample. If a proper random procedure is not strictly adhered to, the investigator's desire or motive to obtain a certain result from the sample survey may influence his selection of the sample, consciously or unconsciously. This type of bias is rather serious, inasmuch as its existence may not be immediately apparent.

Secondly, bias may arise because of *substitution* of a selected unit by another. For instance, in a house to house survey, the next house may be interviewed if there is no reply forthcoming from the selected house. This will introduce substitution bias in the sample and vitiate the result of the inquiry.

Thirdly, bias may arise from *incomplete investigation*, that is, because of not investigating the whole of the chosen sample. In a house to house survey, even if there is no substitution of the selected house by the next house, bias may arise if the selected house is not visited a second time or, if necessary, a third time, to elicit reply from that house. This type of bias arises particularly in the case of postal questionnaires.

Sometimes bias may also arise in the process of *estimation*. For instance, simple average of the recorded yields per acre may be calculated instead of the weighted average of the recorded yields.

Avoidance of Bias.

If possibility of bias exists, the conclusions drawn from the sample cannot be regarded as fully objective. The first essential of any sampling procedure, therefore, is to eliminate all important sources of bias.

To avoid bias in the selection process, the only universally recognised method is to draw the sample either entirely at random or at random subject to such restrictions as, while improving the accuracy, would not introduce bias into the results. Drawing a sample entirely at random does not mean haphazard selection. A random sample can be obtained by

adhering to a scientific random process, such as the use of a table of random numbers or drawing of lots.

Bias arising from substitution or failure to investigate the whole of the selected sample should not be allowed to enter the survey. The sample once selected should be strictly adhered to throughout the inquiry and every attempt should be made to investigate the entire sample. Likewise, bias in estimation should be avoided by using the proper processes of estimation.

The above is not to say that removal of even trivial bias in a sample is absolutely essential. Minor sources of bias are permissible, for the magnitude of errors produced by them would be practically insignificant as compared to the random sampling errors. Of course, avoidance of any substantial bias is of the utmost importance.

Sampling and Non-sampling Errors.

Since a sample survey implies the study of a small proportion of the total universe, there would naturally be a certain amount of inaccuracy in the information collected. These inaccuracies are of two different kinds, one of which is peculiar to sample surveys and the other is common to all surveys including the complete count.

The first kind of error is the sampling error. This arises because the information collected refers to only a part and not the whole. Thus, if two or more samples are taken and further if the required information is collected without any mistakes whatsoever, even then the results obtained from the different samples would usually differ to some extent. These differences constitute the errors due to sampling.

In actual practice, certain inaccuracies and mistakes may also creep in during the process of collecting the actual information. These inaccuracies arise from the negligence or bias of the investigators in asking the questions or in recording the answers. They are also due to the negligence or bias or lack of knowledge or foregetfulness on the part of the informants. These inaccuracies taken together constitute the non-sampling errors and they occur in all surveys—census or sample.

The sampling errors can be estimated mathematically. The non-sampling errors, since they arise from the human factor

and at the stage of the conversation between the investigator and the informant, are not so measurable. It is a fortunate fact that the magnitude of sampling errors can be calculated from the detailed results obtained from an actual sample, for it enables us to make an assessment of the relative accuracy of the different sampling methods employed on the same material. That is, a knowledge of sampling errors resulting from different sampling methods enables us to select the most efficient method of sampling. This advantage of the knowledge of sampling error has changed sampling from a speculative and uncertain procedure to a method having definite and determinable precision. It has added to the reliability of the sampling method. Having known with the help of the sampling error the limits within which the results of other samples drawn from the same universe would be found to lie, we always have an indication of the limits within which the results of a complete census of the same universe, if conducted, would lie. Since this is the virtue of the sampling error, it is obvious that whether or not a sample will give results which are sufficiently representative of the whole universe, would depend primarily on whether the sampling errors are sufficiently small.

Methods of reducing the Random Sampling Error.

In order that the results of a random sample be representative of the entire universe, it should be ensured that the magnitude of random sampling error is small. This can be done firstly by removing all substantial bias from the sample and, secondly, by enlarging the size of the sample. Other things being equal, the sampling error is approximately inversely proportional to the square-root of the number of units included in the sample.

The accuracy of the results of a sample survey depends not only on the number of units included in the sample but also on the variability per unit; or, more correctly on that part of the variability per unit which contributes to the sampling error. This variability per unit can be substantially reduced by imposing certain restrictions on fully random selection. The various devices by which such restrictions are imposed are:

1. Stratification,
2. Multi-stage sampling,

3. Utilization of a variable sampling fraction, and
4. Utilization of supplementary information.

In the first method, the given population may be 'stratified' or divided into blocks or stratum such that the units falling in each block or stratum are as similar or homogeneous as possible. Each of the strata may then be sampled at random. If the same proportion is taken from each stratum, then each stratum will be represented in the correct proportion in the sample and, consequently, differences between different strata will be eliminated from the sampling error.

In multi-stage sampling, the population is divided into a number of first-stage sampling units, from which a sample is selected. The selected first-stage units are then sub-divided into smaller second-stage units from which again a sample is selected. Third-stage sampling may also be done in the same manner, if required. Thus in a rural consumption survey a sample of villages may first be taken and, then, in each of the selected villages a sub-sample of households may be taken. Thereafter, a further sub-sample of individuals may be taken from the selected households, for certain other studies.

Utilisation of a variable sampling fraction implies the inclusion of different proportions of the different strata in the sample, so that the more variable parts of the universe are sampled more intensively. For example, while taking out a sample of workers in a given industry, a much larger fraction may be taken of the larger factories than of the smaller ones. It should, however, be remembered that if the sample is selected in this manner, it will be necessary to weight the contributions of the different strata to the total, in correct proportions.

Supplementary information available from sources other than the sample survey may also be used. For example, in a random sample survey of production of wheat in India, area figures may no doubt be estimated with the help of the survey. The area figures recorded in the village account books of Patwaris may also be used by way of supplementary information, at any rate in areas which are cadastrally surveyed and mapped.

Selecting a Random Sample.

In selecting a random sample, answer has to be found to two questions: (1) How many units of the universe should be included in the sample, and (2) how should these units be selected so that the selection is done objectively?

No hard and fast rule can be laid down in so far as the number of units to be selected from the universe is concerned. The number would depend upon the degree of homogeneity existing in the universe. If the universe were perfectly homogeneous, study of any one unit will give the most accurate results for the entire universe. But universes are hardly ever perfectly homogeneous. In practice, therefore, one may begin with a small sample and increase its size till successive tests indicate sufficiently similar results. When the results obtained from different sizes of samples check up with one another and show the same tendencies, it is clear that the number of units chosen is adequate. As a general rule, when a given proportion of the material is included in the sample, the smaller the sampling units used, the greater will be the accuracy and representativeness of the results. Thus in an agricultural survey it will be more accurate to take, say, 10% of all farms in each *taluk* (*tehsil*) or village than to take all the farms in 10% of the *tehsils* or villages. The reason for this is clear: the farms in any *tehsil* or village are likely to be more alike than those of different *taluks* or villages.

As regards the second question, every unit in the universe should be given an equal chance of being included in the sample so that the results obtained from the sample can be assessed in terms of the Law of Probability. For such objective selection any one of the following methods may be used:

(1) Drawing lots, or lottery sampling.

(2) Arranging the units geographically, numerically or alphabetically and selecting every 10th, or 15th, or 17th item, as required.

(3) Numbering each of the units, preparing one card for each number, shuffling the cards and then drawing the required number of cards at random.

(4) Utilising tables of random numbers.

Tippett's Numbers.

Tables of Random Sampling Numbers have been constructed by L. H. C. Tippett and are known after his name. These tables consist of 41,600 digits taken from census reports and combined by fours to give 10,400 four-figure numbers. The first thirty sets of these numbers are reproduced below:—

2952	6641	3992	9792	7979	5911
3170	5624	4167	9524	1545	1396
7203	5356	1300	2693	2370	7483
3408	2762	3563	1089	6913	7691
0560	5246	1112	6107	6008	8126

Example 1: From the above table of random numbers, required to take a random sample of 12 from 8,000 houses.

All the 8,000 houses are numbered from 1 to 8,000, and we have to find 12 numbers at random between 1 and 8000. From the given table, we select the first 12 numbers which are not greater than 8000. These 12 numbers are:

2952	6641	3992	7979	5911	3170
5624	4167	1545	1396	7203	5356

Conducting a Sample Inquiry.

After the selection of the sample, the next step is to investigate it, or, collect the required data from it. For this purpose, an appropriate method of collection of data will have to be chosen. Several such methods were discussed in Chapter V (*Cf.* pp. 42-45). Like a census inquiry, the reliability of the conclusions emerging from the sample survey would depend on whether the data collected are themselves reliable, which in turn would depend on the method used in the collection of data, the type of investigators employed, the extent of training given to them, the nature of questions put to the informants, and so on. Merely because a survey has been conducted on a random sampling basis, the results following from it do not acquire the hall mark of quality. If the bias of the investigator, the informant or the measuring instrument were allowed to enter, say, a rural consumption enquiry; if the questionnaire adopted were too long or had not been tried and perfected by holding a pilot inquiry beforehand; if the investi-

gators were not given adequate and intensive training or were not conversant with the local language, customs and usage ; if they conducted their inquiry long after the consumption period to which the inquiry related, so that the investigators had no chance of seeing for themselves, much less weighing, what the households concerned really consumed; if the households, knowing that their actual offtake would not be checked up by the investigator, were influenced by a sense of pride to overstate their consumption—in all such cases the results of the inquiry would not be held as reliable, although the households interviewed and the villages in which they were situated were selected on the best random sampling basis possible. It is held in certain quarters that some such defects entered the First Round of the National Sample Survey conducted during October 1950—March 1951 in India. It has, therefore, to be noted that the mere fact that a particular sample has been selected randomly is no guarantee that the conclusions drawn from it must be nothing but the absolute truth. Their worth will also depend on the manner in which the inquiry has been conducted. If the inquiry has been conducted objectively, with all the necessary care and precaution, the fact that the sample was randomly selected would add to the reliability of the data collected as well as to the inferences flowing from them.

Statistical inferences drawn from a sample inquiry have a significance of their own. Whereas such inferences drawn from a census inquiry can be said to be deductions or statistical descriptions, those drawn from a sample inquiry are in the nature of induction inasmuch as they are applied to the entire universe from which the sample was drawn.

The Nature of Statistical Induction.

Induction stands for the logical process by which generalizations are arrived at from a study of particular cases. Statistical induction or statistical inference means the generalization of statistical results, that is, applying to a population the results derived from a sample.

In doing so, the type of problem that arises is really this : suppose we derive a certain measurement, say, an average, or measure of dispersion or correlation, from a sample, then may

we assume that if successive samples were taken from the same population, their measurements too would have the same values as those of the first ? And, if not, may we determine the limits within which the measurements derived from successive samples would fluctuate ? This is the real problem, and one of considerable significance in all sample studies.

In regard to inductive reasoning, two points need to be emphasized. Firstly, the results derived from statistical induction hold only in terms of probabilities. This is so because whereas the results are derived from the study of a part of the universe, they are applied to the whole of the universe ; that is, the results of induction are applied to cases not included in the sample as such. In other words, induction yields results going beyond the premises. Since, in inductive reasoning, we apply to the unknown the results of the known, a leap in the dark is always involved and, in that sense, induction is hazardous.

But, is it really hazardous ? Is there nothing to justify this leap in the dark ? There is, and this constitutes the second point to be emphasized; there is the Law of Statistical Regularity to which nature is generally subject. According to this law, while there is variation in nature, the extent of such variation is rather limited; there is to be seen some uniformity in all natural processes. This uniformity is found in the Inertia of Large Numbers, as exemplified by the strange regularities in such measurements as birth-rates, death-rates, etc. This principle of regularity, or stability, appears in all natural processes and is strongly reflected when huge masses of quantitative data are dealt with. Therefore, when we generalize the conclusions of a sample, we do it on the *assumption* that in the universe to which the conclusions are applied there exists a uniformity in respect of the characteristics we have measured. And, as a result of this uniformity in the universe, we should naturally expect that statistical measurements like measures of central tendency, dispersion or correlation, computed from successive samples drawn from the same universe, would fluctuate within definite limits.

But in this assumption there lies a hypothesis which we cannot verify by purely statistical methods. Accordingly, there is an *a priori* element in every statistical induction. The

statistical conclusion cannot, therefore, stand by itself; it must be corroborated by reason, judgment and whatever internal or external evidences are available.

Statistical Induction and Normal Distribution.

The main task with which we are faced in generalization of the conclusions derived from a sample is that of determining the limits within which the statistical measurements derived from successive samples would probably fluctuate. Here we call to our aid the characteristics of normal distribution or Gaussian curve, discussed earlier. If our universe is assumed to obey the principle of uniformity of nature, that is, if we assume it to be stable, it follows that all its attributes are stable. Under such conditions, the statistical measurements defining the attributes of numerous samples drawn from the same universe will be distributed about the true value (that is, the value of the universe) in a systematic manner. This distribution will closely approximate the normal distribution and, with this knowledge, it is possible for us to determine the limits within which the statistical measures derived from different samples of the same population would lie. A number defining these limits will serve as a measure of the reliability of the given results, when generalised.

Suppose we desire to test the reliability of an average of monthly expenditure of middle-class families in Delhi. A first average might be computed from a sample composed of 200 families. This result might then be tested by computing 399 additional averages, each based on 200 families. These 400 averages would not be identical in value; they would slightly differ from one another. But, if they were tabulated into a frequency distribution, a distribution closely approximating the normal type would result. From this distribution, we may compute the average and the standard deviation. The standard deviation so computed would give us a measure of the variation found in the averages of monthly expenditures, as computed from successive samples.

It follows from the above that the process of random sampling employed in drawing the successive samples should satisfy the conditions giving rise to normal distribution. These conditions are three, as discussed earlier in the section on

Normal Curve of Error, *viz*: multiple causation, independent causation and equality of causal forces (*vide* p. 556). Thus, no special unbalanced influences should affect the successive samples; they should be absolutely random. If so, and if the given universe is homogeneous, the distribution of means derived from the successive samples would be normal. This fact is of great significance to statistical induction. On it rests the entire theory of sampling. Thus the Gaussian law constitutes the base on which is raised the superstructure of statistical induction.

In practice, it is hardly possible, or, even necessary, to draw a large number of successive samples and build a frequency distribution of their averages. In our illustration, it would be impracticable to take 399 successive samples of middle-class families in Delhi to test the reliability of the first average. The illustration only serves to clarify the principle involved. In actual practice, certain processes of mathematical deduction are employed. A study of these processes forms the corner-stone of the theory of sampling.

Theory of Sampling.

While sampling a universe, we usually have in mind one or more of its variates. For example, when we sample the population of India, we are not so much concerned with the individuals as human beings as with one of their qualities such as age or civil condition, or, perhaps with the correlation between age and civil condition. Our object will then be to obtain from the sample an idea of the frequency distribution by age and civil condition in the population of India as a whole. It may perhaps be enough for us to find estimates of certain constants of the universe. Mean and standard deviation may possibly give us a fair idea of the age distribution of the population of India.

Thus a large part of the theory of sampling is concerned with estimating certain constants of the parent universe from the sample. These constants include measures of central tendency, dispersion, skewness and correlation.

There is yet another aim of the theory of sampling. It has already been emphasised that the conclusions derived from a sample are not of the categorical kind generally found in

mathematics, that these conclusions are not based on certainties but on probability. It is, therefore, an important aim of the theory of sampling to determine as objectively as possible what degree of confidence we can put in our statistical measures when they have been obtained from a sample. In other words, our object is to determine the *precision of an estimate.* We attempt to determine the precision of our estimate by assigning limits to the probable divergence between the estimate derived from the sample and the true value of the estimate in the universe. In the science of statistics we deal with both the attributes and variables. Theory of sampling is, therefore, applied to sampling of attributes as well as to sampling of variables.

SIMPLE SAMPLING OF ATTRIBUTES—LARGE SAMPLES

In Chapter XVII on Association of Attributes, the meaning of attributes and variables was explained. The sampling of attributes may be understood as drawing of a sample from a universe containing A's and not-A's. The number of A's in each sample will constitute a part of the data provided by the samples. In keeping with the nomenclature used in the section on Probability, we may speak of the drawing of an individual on sampling as an "event" n, the appearance of attribute A as a "success" p, and the non-appearance as a "failure" q.

Sampling of attributes may be considered as simple and as non-simple, though both may be random. Simple sampling is one in which each event has the same chance of success and in which the chances of success of each event are independent, whether previous trials have been made or not. These conditions hold good in the throwing of a die or the tossing of a coin ; the chance of getting heads with a coin is not affected by the success or failure to obtain it on previous trials, and remains constant irrespective of the number of trials. In contrast with such cases, there may be a case in which we take a sample of two balls from a universe consisting of 8 black balls and 6 white balls under random sampling conditions, that is, such that in each of the two events every member of the universe has an equal chance of being chosen. If in the first

trial we draw a black ball, the chance of doing so will be $\frac{8}{14}$. Then 7 black balls and 6 white balls will be left in the universe, so that the chance of drawing a black ball in the second trial will be $\frac{7}{13}$. This latter chance is not the same as the chance in the first trial and therefore the sampling in this case, though random, is not simple.

Mean and Standard Deviation in Simple Sampling of Attributes.

If we take N samples with n events in each and the chance of success of each event is p and of its failure $q=1-p$, the mean M of the distribution will be given by

$$M=np$$

The standard deviation of the distribution will be given by

$$\sigma=\sqrt{npq} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Instead of recording the number of successes in each sample, if we had recorded the proportion of successes, that is, $\frac{1}{n}$ th of the number in each sample, the standard deviation s of the proportion of successes would be given by

$$s=\sqrt{\frac{pq}{n}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

The above two equations of calculating the standard deviation are of very great importance in sampling and we will now illustrate their use.

Example 2. In Table 68, frequency distribution is given of the observations resulting from 64 throws of 6 dice, a throw of 4, 5 or 6 spot being called a success and that of 1, 2 or 3 spot being called a failure. Required to find the proportion of successes.

If the dice were all true, the chance of success is $\frac{1}{2}$. Hence the theoretical mean M is equal to 3 and the theoretical value of the standard deviation is equal to 1.22.

From the actual observations we had observed that mean

$$M=3.125$$

and, $\sigma=1.1$

The proportion of successes is thus $\frac{3.125}{6}=0.521$, instead of $\frac{1}{2}$ or 0.5.

We have observed in the section on Probability that binomial distribution is a single-humped type and approximates very closely to the normal distribution for large values of n. That is, the great majority of items lie within a range $\pm 3\sigma$ on each side of the mean, or, $\pm 3\sqrt{npq}$ on each side of the value np. If the distribution is exactly normal, 0.9973 of the curve would lie within this range. We may, therefore, conclude that if a particular sample gives a value of p which is outside this range, the deviation from the expected value is most *unlikely* to have arisen from fluctuations of simple sampling, provided, of course, n is large.

Example 3. In an experiment 2,208 throws of a 4, 5 or 6 spot were made out of 4,096 throws altogether, a throw of 4, 5 or 6 being considered a success and that of 1, 2 or 3 spot a failure. The chance of throwing a 4, 5 or 6 spot is $\frac{1}{2}$, and hence the expected value is 2,048. The observed number is thus 160 in excess of the expected number. Can this difference be due to fluctuations of simple sampling ?

The standard deviation of simple sampling is

$$\sigma=\sqrt{npq}=\sqrt{4096\times\tfrac{1}{2}\times\tfrac{1}{2}}=32$$

The deviation of observed number from the expected value is 160, which is 5 times of 32. It is, therefore, most improbable that this deviation arose as a sampling fluctuation.

Example 4. If in example 3, the throws of a 4, 5 or 6 spot numbered 2,064, the observed number would exceed the expected value by 16 which is only 0.5 of the standard deviation of simple sampling. The deviation, in that case, could well have arisen from fluctuations of simple sampling.

Standard Error.

The standard deviation of simple sampling calculated in examples 3 and 4 above, defines the sampling error, and is given a briefer name—Standard Error. The use of the word error may be justified in this connection on the ground that the expected frequency is taken as the true value and divergences of observed frequency from the expected frequency are

taken as errors of estimation due to sampling effects. As a working rule, it should be remembered that if observed frequencies differ from the expected frequency by more than 3 times the standard error, then such divergence is almost certainly not due to fluctuations of simple sampling; it indicates some departure of the sampling from simplicity, which may arise from a flaw in the sampling technique or from causal effects in the population.

In examples 3 and 4, p (that is, the probability of success) was known *a priori*. In many cases it is not so known. In such cases the proportion of successes in the sample is assumed to be an estimate of the proportion of successes in the population. This assumption is justifiable if n is large and neither p nor q is small.

Example 5. From the weather records of a certain district a sample of 800 days is drawn and 100 of them are found to be rainy days. What are the probable limits to the percentage of rainy days in the district ?

Assuming that the conditions of a simple sample hold good in this case, we have

$$p=\tfrac{1}{8} \qquad q=\tfrac{7}{8}$$

Standard error of the proportion of rainy days $=\sqrt{\dfrac{pq}{n}}$

$$=\sqrt{\frac{1}{8}\times\frac{7}{8}\times\frac{1}{800}}$$

$$=0.0117$$

$$=1.17 \text{ per cent.}$$

Therefore, taking $\tfrac{1}{8}$ (or, 12.5 per cent) to be the estimate of the number of rainy days, the probable limits to the percentage of rainy days in the district concerned are 12.5 per cent$\pm(3\times1.17)$ per cent, that is, 9 per cent and 16 per cent approximately.

It will be seen from the above that the value of standard error depends entirely on the value of p and the size of the sample. Therefore, the range within which p probably lies is independent of the size of the population. That is why the size of the population does not appear in the formula for the standard error.

It is significant to remember that the standard error of p varies inversely as the square root of n, and not inversely as n itself. Therefore, as n becomes larger the standard error becomes smaller, the standard error decreasing proportionately to the square root of n. For example, if a sample of 100 gives a standard error of 10%, it will take a sample of 400 to halve that error and a sample of 10,000 to reduce it to 1%.

Precision.

The standard error measures the *unreliability* of the value of p. The greater the standard error, the greater are the fluctuations of the observed proportion. The reciprocal of the standard error ($\frac{1}{s}$), on the other hand, is a measure of *reliability,* also called precision. Therefore, the precision of an observed proportion varies as the square root of the number of observations. To double the precision, the number of observations should be increased four times.

Limitations of Simple Sampling.

We may turn to the formulæ of equations (1) and (2), on page 573, and pose the question: Under what conditions should we expect the observed frequencies to show a standard deviation given by equation (2)? To answer this query, it is necessary to consider what the conditions are which give rise to simple sampling in practice. These conditions are three, *viz.,*

(1) The probability p of drawing an individual with attribute A must remain constant and be the same for all samples. This would happen only when the proportion of individuals with attribute A in the population remains constant at the drawing of each sample. Thus, formula (2) will hold good in case of sampling of persons possessing certain characteristics—blue eyes or dark hair or blindness—if there is no difference in any essential respect between the localities from which the samples are taken or during the period over which the observations are spread.

(2) The proportion of individuals with attribute A must remain constant at the drawing of each individual member of the sample.

(3) The individual events must be independent of one another, like the throws of a die.

Thus, in a sample inquiry relating to the number of persons dying in a series of 1,000 persons (*i.e.*, an inquiry into death rate), condition (1) would not be satisfied if the localities selected constituted a mixture of healthy and unhealthy districts or contained population of different age and sex composition, or if the observations were spread over a period during which sanitary conditions were continuously improving; condition (2) would not be satisfied unless each sample included persons of one sex and one age; and, condition (3) would not be satisfied if we were dealing with deaths from, say, an infectious or contagious disease, for if such disease were contracted by one person the probability of others falling victim to it would be increased. These conditions do not leave the room wide for application of the formulæ in question to cases of economic and sociological character. The formulæ, however, apply, to a large extent, to certain biological cases.

SIMPLE SAMPLING OF VARIABLES—LARGE SAMPLES

In the sampling of variables we are no longer concerned with classifying each member of a sample under one of the two heads, success or failure. We are concerned here with frequency distributions, such as agricultural wages, human consumption, industrial production or rural indebtedness. But, our objects of studying the samples of variables are the same as those of studying the samples of attributes, *viz.* (i) to compare observation with expectation and to ascertain how far the divergence of one from the other can be attributed to fluctuations of sampling, (ii) to estimate some characteristic, say mean, of the universe from the sample and (iii) to assess the reliability of such estimate.

Sampling Distribution.

If we take several samples from a population and calculate some function,* such as the mean, the standard deviation or the coefficient of correlation, of each sample, it is not necessary

* Measurements such as mean, standard deviation, coefficient of correlation, etc., are usually referred to generically as parameters.

that the values of such function for all the samples should be identical. In fact, we will get a series of different values. If the number of samples is large, we may group these values in a frequency distribution. As the number of samples increases. this frequency distribution will approach the normal curve. Such a distribution is designated a Sampling Distribution.

The chief characteristic of sampling distribution is that it has a peak somewhere near the value of the parameter which corresponds to the true value in the parent population. That is, a large majority of the values of the parameter in question cluster round the peak and only a few are scattered far from the true value of the parameter in the universe. If this characteristic is not satisfied by a given sampling distribution, the distribution would probably be a biased one.

Simple Sampling of Variables.

The discussion on simple sampling of attributes applies to simple sampling of variables too. The conditions which give rise to simple sampling of attributes and which had been explained on p. 576 under "Limitations of Simple Sampling" also apply to the simple sampling of variables. Thus, in simple sampling of variables we assume that (i) we are drawing our samples from the same record, (ii) each member of our sample at each drawing is drawn from the same record, and (iii) the drawing of each member of our sample is entirely independent of that of every other. If these conditions are satisfied, the sampling distribution can be represented satisfactorily by a mean and standard deviation which may be estimated from the data of the sample, as in the case of simple sampling of attributes. We can take a range of ± 3 times the standard error as giving the limits outside which the value of the parameter measured from a sample would probably not lie. We can thus use the standard error, as we did in the case of sampling of attributes, to determine the precision of an estimate or to give a judgment on the divergence between the expected and observed values.

We may now take up a discussion of the standard errors of different statistical measurements.

Standard Error of Arithmetic Mean.

The formula for the standard error of arithmetic mean, σ_M, is

$$\sigma_M = \frac{\sigma}{\sqrt{n}}$$

where σ in the right-hand member of the equation stands for the standard deviation of the population from which the sample is taken and n for the number of observations in the sample. Actually, σ is not known ; the standard deviation of the sample is therefore, used as an approximation to it. This approximation is permissible when n is large, and the above formula for σ_M becomes

$$\sigma_M = \frac{s}{\sqrt{n-1}}$$

where s is the standard deviation of the sample, and has been derived from the relation

$$s = \sqrt{\frac{\Sigma d^2}{n}}$$

where d is the deviation of a single observation from the mean and n is the number of observations.

What use do we make of the knowledge of the standard error of the mean? We have noted that a sampling distribution of arithmetic means (that is, a distribution of means derived from random samples of a given population) usually follows the normal law of error. This is true even if the distribution of the population from which the samples are drawn is not itself normal. Accordingly, it is possible for us to interpret given values of σ_M with reference to the probabilities associated with deviations in a normal distribution. That is, it is possible for us to set the limits within which it is highly likely that the true mean—the mean of the population—would lie.

Suppose in a given case the sample mean is 50 and the standard error 0.18. We might, then, make the statement that the true mean, which is unknown, lies between 50 ± 0.18, that is, between 49.82 and 50.18. We may not dogmatise on this statement ; but, if we take an infinitely large number of successive samples from the same population computing the mean of each sample and the standard error of each mean, the

chances are 68 out of 100 that our statement will come true.

This statement would vary with the degree of probability we want. We might, for instance, make the statement that the true mean lies between $50 \pm (1.96 \times 0.18)$, that is, between 49.65 and 50.35. If we take an infinitely large number of similar samples, our statement would be true in 95 cases out of 100. (We have used the relation $M \pm 1.96\sigma$ in this case on the basis that 95% of the area under normal curve lies within $M \pm 1.96\sigma$. In usual practice, the relation $M \pm 2\sigma$ is used).

From the above it will be seen that one use of the standard error is to measure sampling reliability. Another use of this error is to test a given hypothesis, that is, to test whether a given sample could have been drawn from the parent population. This test involves the use of tables of normal probability integral. As a guiding principle, however, it will be useful to remember the Test of Significance described below.

Test of Significance.

In using the standard error as an instrument for testing a given hypothesis, the hypothesis usually involves an assumed parent population and the test only purports to assess whether a given sample could have been taken from this parent population.

Suppose on rational grounds we establish the hypothesis that the average monthly expenditure of high-school students in the city of Allahabad is Rs. 50. We then conduct a sample survey of the monthly expenditure of 226 high-school students in that city and the average expenditure is found to be Rs. 49.5 with a standard deviation of Rs. 1.8. The standard error of the mean is

$$\sigma_M = \frac{s}{\sqrt{n-1}}$$

$$= \frac{1.8}{\sqrt{226-1}}$$

$$= \frac{1.8}{15}$$

$$= 0.12$$

The question that now arises is as to whether these results are consistent with the hypothesis that our sample of 226 students is drawn from a parent population with a mean expenditure of Rs. 50 per month. The formula for this test to use would be

$$T = \frac{M - M_h}{\sigma_M}$$

where T stands for the deviation of sample mean from the hypothetical mean, expressed in units of standard error of the mean ; M for the mean of the sample ; M_h for the hypothetical mean of the population ; and, σ_M for the standard error of the mean.

As a general working rule it may be remembered that if T gives a value of 2.576 or more, the difference it measures is significant ; and, if T gives a value of less than 2.576 the difference is not significant, that is, the results are not inconsistent with the hypothesis.

Example 6. In the case of the average monthly expenditure of high-school students for which the data are given above, required to say whether the results are consistent with the hypothesis.

Now,

$$T = \frac{M - M_h}{\sigma_M}$$

$$= \frac{49.5 - 50}{0.12}$$

$$= \frac{-0.5}{0.12}$$

$$= -4.16$$

The sample mean thus represents a departure of 4.16 standard deviations from the general mean. If this departure were 2.576 we could say that the results were consistent. Since that is not the case, we conclude that the results are not consistent with the hypothesis. The difference is significant, that is, not due to chance. In other words, we could not have secured the sample mean of Rs. 49.5 and standard deviation of Rs. 1.8, had we been drawing from a population of high-school students with a mean expenditure of Rs. 50 per month.

Example 7. In the above case of high-school students, let us assume that the hypothesis is that the mean expenditure is Rs. 49.45 and then test if the results are consistent.

Now,

$$T = \frac{M - M}{\sigma_M}$$

$$= \frac{49.5 - 49.45}{0.12}$$

$$= 0.416$$

The deviation in this case is much lower than 2.576. It is insignificant. Such a deviation is highly probable. The results are, therefore, not inconsistent with the hypothesis that the mean expenditure is Rs. 49.45 per month.

The results in example 7 do not, however, prove the hypothesis. The tenability of the hypothesis must be assessed on the basis of rational considerations and empirical evidence and on the knowledge we have regarding the parent population.

Sampling Error: Median and Quartiles.

The median is subject to greater fluctuations of sampling than the mean. The degree of dispersion of medians derived from successive samples is about 25% greater than that of arithmetic means of the same samples, so that,

$$\sigma_{Md} = 1.25331 \frac{s}{\sqrt{n-1}}$$

Where σ_{Md} is the standard error of the median. The formula for the standard error of the quartiles is

$$\sigma_{1} = \sigma_{Q3} = 1.36263 \frac{s}{\sqrt{n-1}}$$

Sampling Errors: Standard Deviation.

In the case of a normal distribution of the universe, the standard error of standard deviation σ_σ is given by

$$\sigma_\sigma = \frac{s}{\sqrt{2n}}$$

Sampling Errors: Coefficient of Correlation.

The formula for the sampling error of the coefficient of correlation, σ_r, computed from n pairs of observation is given by

$$\sigma_r = \frac{1-r^2}{\sqrt{n-1}}$$

where r in the numerator of the right-hand member is the true coefficient of correlation in the population at large. Since the true r is not known, r of the sample is used instead.

The above formula also holds for distributions approaching the normal type, when n is fairly large.

The test of significance in the case of coefficient of correlation is given by the formula :

$$T = \frac{r - r_h}{\sigma_r}$$

where r is the coefficient of correlation of the sample and r_h the hypothetical coefficient of correlation.

How probable error helps in determining the significance of the value of r has already been discussed in Chapter XV, pp. 429-431, and repetition of the same here is considered unnecessary.

Sampling Error: Coefficient of Regression.

The formula for the standard error of coefficient of regression is given by

$$\sigma_b = \frac{s_y}{\sqrt{\Sigma x^2}}$$

where σ_b stands for standard error of regression, x for the given value of the independent variable expressed as a deviation from the mean of that variable, and s_y for the root mean square of the deviations of the actual values of y, the dependent variable, from the corresponding computed values.

The test of significance in this case is given by

$$T = \frac{b - \beta}{\sigma_b}$$

where b stands for the coefficient of regression of the sample and β for the hypothetical coefficient of regression.

Sampling Errors: Difference between Means.

If two samples are drawn from the same universe there will naturally be some difference between their means. The problem as to whether the difference between two such means is significant, therefore, arises frequently in sample studies. The problem, in other words, is to determine whether the two samples in question could have been drawn from the same population.

The procedure employed in determining the significance of a difference between two sample means is, like any other such measure, dependent on the normal law. If a number of *pairs* of samples were drawn from a given population, the difference D between the two means of each pair could be computed from M_1-M_2. A frequency distribution of the differences thus secured (that is, of D's) would follow the normal law. The standard deviation of this distribution of D's is given by

$$\sigma_D = \frac{\sigma_1^2}{n_1-1}+\frac{\sigma_2^2}{n_2-1}$$

which comes to the same thing as

$$\sigma_D = \sqrt{\sigma_{M1}^2+\sigma_{M2}^2}$$

The test of significance for the difference between the two means is given by

$$T=\frac{D-D_h}{\sigma_D}$$

where D is difference between the two sample means, D_h the hypothetical difference which is taken as zero and σ_D the standard deviation of the distribution of D's. If, as in the tests of significance referred to earlier, T exceeds a value of 2.576, the difference between the two means is significant, otherwise not.

Example 8. Given that in a sample of 200 villages the mean population was found to be 485 persons with a standard deviation of 50 and in another sample of the same number of villages the mean and standard deviations were, respectively, 510 and 40. Required to find whether the difference between the two means is significant.

We start with the null hypothesis: the averages of the two samples are not significantly different. Then, if the value of

T comes out to be less than 2.576 (or, in other words, lies within ± 2.576), our hypothesis will stand, otherwise not. Now, the standard error of the difference between the two means is

$$\sigma_D = \sqrt{\frac{\sigma_1^2}{n_1-1} + \frac{\sigma_2^2}{n_2-1}}$$
$$= \sqrt{\frac{(50)^2}{200-1} + \frac{(40)^2}{200-1}}$$
$$= \sqrt{20.6}$$
$$= 4.539$$

The value of D is $(485-510) = -25$, or 25, for the algebraic sign does not matter. This value of D is to be judged with reference to a hypothetical value of zero. Accordingly for T, the discrepancy expressed as a normal deviate, we have

$$\mathrm{T} = \frac{D-0}{4.539}$$
$$= \frac{25}{4.539}$$
$$= 5.5$$

This discrepancy is much larger than the magnitude 2.576. Our null hypothesis does not stand. The difference between the two means is, therefore, significant; it is not due to chance; it is not due to fluctuations of sampling. The magnitude of discrepancy is not consistent with the hypothesis that the two samples were drawn from the same population.

Value of T.

A working rule has been stated above that if the value of T is 2.576 or more, it indicates a significant difference, whereas if the value of T is less than 2.576, it indicates that the results are not inconsistent with the hypothesis. What is the basis for this working rule? What, in other words, is the reasonable criterion as to the limits of significance? One conventional standard is based on odds of 1 out of 100; if a given difference between hypothetical estimate and observed value would occur as a result of chance only 1 time out of 100, or even less frequently, we may call the difference *significant*; that is, the

results are not consistent with the hypothesis. If on the other hand, a given difference occurs more frequently than 1 time out of 100 as a result of pure chance, we may call the difference as *not clearly significant;* that is, the results are not inconsistent with the hypothesis. These two criteria constitute the basis of our working rule. *T*, as we know, is the difference between the hypothetical value and the observed value of a statistical measurement, in units of the standard error of that measurement. The value of *T* corresponding to a probability of 1 out of 100 is 2.576; one-100th part of the area under the normal curve lies at a distance of 2.576 standard deviations or more from the mean, on the x-axis. As a rough and ready measure, 2.576 is approximated to 3 and 3 is used in place of 2.576.

Some statisticians use odds of 1 out of 20, or, 5 out of 100 as the limits of significance, rather than odds of 1 out of 100. In that case, a difference between hypothetical and observed values that would occur only 5 times out of 100, or, less frequently, would be taken as significant. The value of *T* corresponding to this standard is 1.96 (or, roughly 2). Therefore, a value for *T* of 1.96 (or, roughly 2) or more would indicate a significant difference, while a value of less than 1.96 (or, less than 2) would indicate that the results are not inconsistent with the hypothesis.

Limitations of Measures of Sampling Errors.

The measures of reliability, discussed so far, are of great importance in sampling inasmuch as we may use them to give precision to our judgment regarding the margins of error involved in extending statistical results of a sample to the population. Yet, these measures are subject to certain limitations.

The assumption underlying the interpretation of the measures of sampling errors is that statistical measurements (like means, standard deviations, coefficients of correlation) obtained from successive samples are distributed in accordance with the normal law of error. This assumption is justified only when the number of cases is large, even though the population from which the samples are drawn may not itself be distributed in accordance with that law. If, however, the number

of cases in each sample is small, the assumption does not stand. Therefore, the significance of a given difference in terms of T is materially altered if we are dealing with small samples. Techniques developed for defining sampling errors based on small samples have not been discussed in this book.

It will also be remembered that the standard errors, discussed earlier, measure only those errors which arise from the fluctuations of simple sampling. Therefore, our data must satisfy the three conditions of simple sampling, already discussed (*vide* p. 576). But these conditions of simple sampling are hardly fulfilled perfectly in the field of economics. Conditions of production, trade, income distribution, etc., change from year to year, and even from month to month. It is difficult to assume that data relating to different times reflect the play of identical forces. The standard errors, described earlier, can give no indication of the possibility of fluctuations in successive samples due to causes other than fluctuations of simple sampling. Fluctuations arising from bias, from lack of representativeness in the sample or from any other non-sampling error quite elude the method of determining reliability through standard errors. All the same, the method has its own usefulness.

General Remarks.

The importance of sample surveys in the field of economic and social inquiries has been stressed in the foregoing pages. The place which complete enumeration occupies in such inquiries has also been explained. Generally speaking, random sample survey has two advantages over complete enumeration and is preferred to the latter because of those advantages. Firstly, the method of random sample survey is objective in its approach and, secondly, the method enables us to determine the margin of error of the estimate resulting from the survey.

A sample should be randomly selected, if it has to claim the above-mentioned two advantages. Then alone will it serve as a representative of the population from which it is drawn, that is, then alone will the results derived from it be applicable to the population. To make the sample representative of the population, it is necessary that bias in selection should be avoided and this can be done in many ways. Generally, the

method of stratified random sampling is adopted. In certain cases, inter-penetrating samples are taken so that, as in the case of area sampling, one sample penetrates into another and two independent sets of investigators survey the area common to the two inter-penetrating samples. This acts as a check on the reliability of the survey. In some other cases, cluster sampling may also be adopted as a means of adding to the reliability of the survey. Thus, in a survey of fields or villages, a cluster of fields or villages may be built around the randomly selected field or village. The first major step in sampling, therefore, is to exercise all precautions to see that the sample is objectively selected and that no subjective bias is allowed to enter the selection of the sample.

The next important step is to decide on the method of surveying the sample. The choice of the method of collection of data will have much to do with the reliability of the results. Merely because the sample has been randomly selected, its results may not be valid for the entire population unless due care and caution have been exercised in so far as the investigation itself is concerned.

Sampling is a process of induction and, therefore, its results may be tested with reference to certain internal and external checks and, also, with reference to the sampling errors. In the case of a consumption survey, for instance, it has to be seen as to whether the different results of the survey are internally consistent, whether one finding does not contradict the other, or, whether the different findings are in harmony with one another. If the amount of consumer-expenditure is found in all cases to exceed the amount of consumer-earnings, internal consistency of the survey is open to some doubt. Similarly, by way of external check, the total consumption of a particular commodity may be compared with, say, the total production or total availability of that commodity, if known through some other source. If the total human consumption is found to exceed the total availability for human consumption, then, assuming that the figures leading to total availability are correct, the consumption figure would not satisfy the test of external check. Likewise, when the results of a sample survey are examined, a knowledge of sampling errors is absolutely essential, for, it is these errors which determine the limits

within which it is probable that the true value would lie. If the magnitude of sampling error is big, that would throw up doubts about the applicability of the results of the sample survey to the population.

We have dealt with conditions of simple sampling and with large samples in this Chapter. We have left small samples out. The chief merit of sampling lies in the fact that the results obtained from it can be assessed in terms of probability. The theory of probability, therefore, constitutes the very foundation of the sampling theory. We have given some idea of the principles underlying the theory of probability in this Chapter. The main point to bear in mind is about the characteristics of the normal law of error. We have seen what important role is played by the normal curve of error in this inductive method of study. If our sample fails to satisfy the conditions attending the normal law of error, we cannot put our faith in the results shown by it. Sampling is, therefore, as delicate as it is scientific. It is, therefore, all the more necessary that sample surveys are entrusted to experts in the science, who are conversant with not only the technique but also with the general characteristics of the particular phenomenon they are called upon to study.

EXERCISES

(1) What is meant by the normal curve of error? What are its characteristics? How do they help in assessing the results of sample inquiries?

(2) If we take 100 sets of 10 tosses of an unbiased coin, in how many cases should we expect to get 7 heads and 3 tails?

(3) In exercise (2) above, in how many cases should we expect to get 7 heads *at least?*

(4) (*a*) What is the chance of securing the total 6 from two dice thrown simultaneously?

(*b*) What is the chance of securing *at least* 6 in the toss of two dice?

(5) Briefly describe the method of random sampling, and comment on its uses in social investigations.

(M. Com., Agra, 1945).

(6) Explain the term 'population' and 'random sample'. A random sample of 20 paddy fields was chosen from a district.

The yields of paddy, in maunds per acre, in the twenty fields were as under:—

85.2, 104.3, 97.8, 125.3, 164.3, 143.2, 94.6, 109.3, 111.4, 115.8, 132.6, 151.4, 100.0, 118.2, 97.0, 88.8, 87.6, 75.2, 121.0, 120.0.

What conclusions can you draw about the average yield of paddy per acre in the district?

(M.A., Agra, 1945).

(7) Show the necessity of the use of the method of random sampling in any extensive investigation. How would you make use of this method in carrying out an economic survey of the rural areas of C. P.?

(B. Com., Nagpur, 1945).

(8) What is meant by "sample" method of inquiry? When is it adopted and what are its advantages? Describe the test that may be applied to determine whether the sample is representative or not.

(B. Com., Hons., Andhra, 1944).

(9) What do you understand by 'random sample'? Is it a synonym for 'representative sample'? Why is a 'random sample' supposed to speak for the 'population'? To what types of enquiries is the technique of random sampling specially applicable?

(M.A., Raj., 1951).

(10) 'Random sampling owes its importance to the fact that we can assess the results obtained from it in terms of probability'.

Elucidate this statement, and also discuss the technique of random sampling in this connection.

(M.A., Alld., 1950).

(11) Discuss the relative advantages and disadvantages of the method of complete enumeration and the method of random sample survey in social and economic enquiries.

(M.A., Punjab, Sept., 1952).

(12) 'One of the aims of statistics is to describe population (through sample), and to this end statistical constants are calculated'.

Discuss fully the above statement and show how this is sought to be achieved.

(M.A., Patna, 1940).

(13) Compare the relative advantages and disadvantages of the method of "complete enumeration" and the "method of random sample survey". Explain with reasons the method you will adopt in enquiries relating to (a) area under rice and (b) cost of production of sugar cane in Behar.

(M.A., Patna, 1942).

(14) Suggest the possible sources of bias in the following:—

(*a*) A heap of mangoes is sampled by taking a handful from the top.

(*b*) In drawing a sample of telephone-owners, the Telephone Directory is opened at random and telephone numbers appearing on each left hand page opened in each trial are noted.

(*c*) Readers of a movie magazine are sampled by printing in it an invitation to them to send up their views on a particular film.

(*d*) Investigation into household expenditure in a town is conducted from house to house, ignoring those houses at which the householders are not available.

(15) Weldon threw 12 dice 4096 times, a throw of 4, 5 or 6 being called a success. The following was the frequency distribution observed:

Success	Frequency	Success	Frequency
0	0	7	847
1	7	8	536
2	60	9	257
3	198	10	71
4	430	11	11
5	731	12	0
6	948		
		Total	4,096

Compare the observed frequencies with the expected frequencies. Also compare the actual mean and standard deviation with those of the expected distribution.

(16) A coin is tossed 1,600 times and heads appear 830 times. Is the coin biased? Discuss the theoretical principles involved.

(B. Com., Hons., Andhra, 1943).

(17) In a dice-throwing experiment 25,145 throws of a 4, 5 or 6 spot were made out of a total of 49,152 throws. Each 4, 5 or 6 spot was considered a success. Can the deviation of observed frequency from the expected value be due to fluctuations of sampling?

(18) A biased rupee coin is tossed 100 times and comes down heads 70 times. What are the probable limits to the probability of getting a head in a single trial.

(19) 400 oranges are taken at random from a large consignment and 50 are found to be bad. Estimate the percentage of bad oranges in the consignment and assign the limits within which the percentage probably lies.

(20) Given the following frequencies in an association table:—

$(AB) = 309$ $\quad$ $(aB) = 132$

$(A\beta) = 214$ $\quad$ $(a\beta) = 119$

Do you think that the association could have arisen as a fluctuation of simple sampling, the true association being zero?

(21) (*a*) A bag contains 15 white and 17 black balls. If one draw of 2 balls is made, what is the probability that one of them is white and the other black?

(*b*) One purse contains 1 sovereign and 4 rupees, a second one contains 2 sovereigns and 3 rupees, and a third one contains 3 sovereigns and 2 rupees. One purse is selected at random out of these and a coin is taken out from it. What is the probability that it is a sovereign?

(M. Com., Alld., 1952).

(22) Five coins are tossed 3,200 times and the following results are obtained:—

No. of heads		
0	..	80
1	..	570
2	..	1,100
3	..	900
4	..	500
5	..	50
		3,200

Test the hypothesis that the coins are unbiassed. You can make use of the following data in drawing your conclusion:—

Degrees of freedom		Chi-square value at 5% level of significance
1	..	3.841
2	..	5.991
3	..	7.815
4	..	9.488
5	..	11.070
6	..	12.592

(M. Com., Alld., 1952).

(23) In a random sample of 500 persons from Madras, 200 are found to be consumers of vegetable oil. In another sample of 400 persons from Calcutta, 200 are found to be consumers of vegetable oil. Discuss whether the data reveal a significant difference between Madras and Calcutta so far as the proportion of vegetable oil consumers is concerned.

(24) In a newspaper editorial of 1600 words in Hindi, 36 per cent of the words are of Persian origin. Assuming that simple sampling conditions hold, estimate the proportion of Persian words in the writer's vocabulary and assign limits to that proportion.

(25) From the table given below, calculate the standard errors the median and two quartiles:

Frequency Distribution of Marks in Economics secured at the High-School examination.

Marks		No. of Students
0— 5	..	50
5—10	..	351
10—15	..	1,502
15—20	..	3,000
20—25	..	5,373
25—30	..	4,960
30—35	..	3,032
35—40	..	2,000
40—45	..	981
45—50	..	30
	Total ..	21,279

[The data are imaginary].

(26) From the table given in exercise 25 above, find the standard error of the mean and compare it with the standard error of the median.

(27) A random sample of 1000 agricultural workers in north Bihar shows their mean earnings to be Rs. 2-7-0 per day, with a standard deviation of Rs. 1-8-0. Another sample of 1500 workers in South Bihar gives a mean earning of Rs. 2-9-0 with a standard deviation of Rs. 2-0-0. Discuss the suggestion that the mean rate of earnings varies as between the two regions of Bihar.

(28) A random sample of 1000 farms in 1949-50 in Punjab gives an average yield of wheat of 800 lbs. per acre with a standard deviation of 62 lbs. Another random sample of 1000 farms in the same year gives an average yield of 760 lbs. with a standard deviation of 50 lbs. Is the difference between average yields significant?

(29) A random sample of 200 villages was taken from Agra district and average population per village was found to be 485 with a standard deviation of 50. Another random sample of 200 villages from the same district gave an average

population of 510 per village with a standard deviation of 40. Is the difference between the averages of the two samples statistically significant? Give reasons.

(M. Com., Raj., 1951).

(30) The data below give the frequency distribution of the weights of 629 Indian soldiers. Calculate the average weight of a soldier and its standard error.

Weight in lbs.		Frequency
105—	..	15
116—	..	43
127—	..	138
138—	..	162
149—	..	129
160—	..	82
171—	..	35
182—	..	16
193—	..	5
204—	..	3
215—226	..	1

(M.A., Raj., 1950).

(31) Explain what do you understand by 'Standard Error' of the mean of a random sample.

The data concerning height measurements for a random sample of individuals from a given population are as follows:—

Mean = 172 cm.

S.D. = 12 cm.

N = 65

If a large number of samples of the same size were selected at random from the given population, what would be the limits of the 2% confidence interval for the true mean?

(M. Com., Alld., 1946).

(32) Calculate the standard error from the following data collected in one of the many random sample enquiries conducted to find out the average earnings of a particular class:—

Earnings per month in rupees.	Number of persons.
Up to 10	50
,, 20	150
,, 30	300
,, 40	500
,, 50	700
,, 60	800
,, 70	900
,, 80	1000

(M. Com., Alld., 1951).

(33) Explain what is meant by the standard error of a mean and find its value for the marks in English shown in the table below:

Frequency Distribution of Marks in English.

Range	No. of candidates
0—10	33
11—20	108
21—30	221
31—40	439
41—50	526
51—60	495
61—70	322
71—80	153
81—90	53
91–100	7

The mean mark in Sanskrit for the same candidates was 39.78 and the standard deviation was 14.21. Discuss whether there was any significant difference in the mean marks in the two subjects.

(34) Calculate the standard errors of (*a*) Lower Quartile, (*b*) Median, (*c*) Upper Quartile, and (*d*) Mean for the

distribution of widows by age-groups in India (undivided) as per table given below:—

Age	Per 1000 in 1931	Age	Per 1000 in 1931
All ages	155	20—30	78
0— 5	1	30—40	212
5—10	5	40—60	507
10—15	10	60 & above	802
15—20	34		

(35) (*a*) Explain the concept of standard error and its use in statistics.

A coin is tossed 10,000 times and it turns with head upwards 5,195 times. Will it be reasonable to think that the coin is unbiassed?

(*b*) Intelligence tests on two groups of students, A and B, give the following results:

A : Mean 84 ; S.D. 10 ; number 121.
B : Mean 81 ; S.D. 12 ; number 81.

Explain if the difference is significant.

(36) Below are given the rates of yield of rice in maunds per acre in 500 plots.

Maunds per acre	Frequency
6.75—	3
7.25—	16
7.75—	19
8.25—	48
8.75—	62
9.25—	79
9.75—	88
10.25—	68
10.75—	60
11.25—	34
11.75—	11
12.25—	9
12.75—13.25	3

Explain how standard error can be used to test whether the mean yield of rice can be considered to differ significantly from a standard rate of yield of 10 maunds per acre.

CHAPTER XXI

INTERPRETATION OF DATA

The Science of Statistics, as we have defined it in Chapter II[1], is concerned with the study of the methods of collection, analysis and interpretation of quantitative data. So far, we have been dealing with the various statistical methods employed in the collection and analysis of numerical facts. It is proposed to deal now with the interpretation of statistical data.

Interpretation.

Interpretation stands for the technique of drawing inferences from an analytical study of the collected figures. While discussing the methods of collection, presentation, comparison, correlation, interpolation, regression, business forecasting, etc., we have put down, wherever necessary, in appropriate places the essential precautions which must be kept in view in analysing statistical data as well as commenting on the results of analysis. A repetition of all those cautions here is obviously uncalled for. But, it must be said that commonsense is as much a chief requisite and experience as great a teacher in the delicate task of interpretation as in collection and analysis of quantitative data. If this fundamental principle is ignored, fallacious inferences would be drawn, which would recoil on the statistician and his science. The statistician, to repeat what we said in Chapter II[2], is not an alchemist expected to produce gold from any worthless material ; he is rather like a chemist capable of assaying the value the material contains and of extracting nothing more than this value. Freedom from bias and prejudice on the part of the statistician is, no doubt, necessary in collection and analysis of data ; it is all the more so in interpretation, because it is interpretation more than— or, rather than—collection and analysis of data with which the layman is concerned. The *power* which figures carry with them is such that the layman can be as easily *impressed* by them as he can be *deceived*. From the advertiser's gallary,

1-2 *Cf.* pages 18-19.

from the electioneering platform, from the propagandist's forum, from the partisan press and from a hundred other sources the man in the street is bombarded with tendentious figures put forward to support some *ex parte* statement .Sometimes such figures are reasonably and justifiably used to form a basis for the arguments built upon them ; more often, they give an exaggerated picture of the truth, which may be due to ignorance or inadvertence, but has also been found to be influenced by motive, by deliberate wish to mislead. The layman is not unaware of this fact. If he distrusts all arguments based on figures, his attitude is like that of a reasonable man, who has not the training for himself to separate the wheat from the chaff, and is, therefore, inclined to suspect everything. Statistical methods are most dangerous tools in the hands of the inexpert. A taxi driven by one who does not know the art of driving might fall in a ditch or collide against other vehicles resulting in serious injury to the passers-by, the taxi and the driver. A taxi must, therefore, be driven by one expert in driving. So must statistics be handled by one who is expert in the subject. Most often what happens is that, attracted by the power of impression which statistics command, so many men are led to use them without knowing their limitations, and feel jubilant if they win their point. At other times, it happens that when the data have been scientifically collected and analyzed, they fall into the hands of such people for purposes of interpretation as hardly know the subject. These people, sometimes under the impress of their preconceived notions, and at other times due to their habit of criticising every thing they come across, indulge in hair-splitting, as if they alone know the art of interpreting statistical data. It is, therefore, established that if the task of interpretation is to be scientifically performed, it must be done by a true statistician who is above all prejudice and has the daring to call a spade a spade.

Preliminaries to Interpretation.

Before starting on interpretation the statistician should examine whether

(1) the data are adequate to base his judgment upon ;
(2) the data are homogeneous and comparable ;

(3) the data have been properly collected and are without biased errors ;

(4) the data have been scientifically analyzed, and all disturbing factors considered.

After satisfying himself on these preliminary points he should draw out reasonable inferences. Most of the mistakes that are made in interpreting figurative data arise from false generalisation, a few examples of which may now be considered.

Mistakes due to False Generalisation.

Let us suppose that an argument runs as follows :

The prices of agricultural commodities in India in 1943 were five times the prices in 1931. India's prosperity in 1943 had, consequently, increased by leaps and bounds.

The argument as it stands looks sound. Let it be agreed that the ratio of prices between the two year is correctly stated. Now, 1931 was a year of depression, when agricultural commodities were, indeed, selling at very low prices, and 1943 a year of war boom. Therefore, the two periods are different, and allowance must be made for this difference before right comparison is possible. Are prices alone a measure of prosperity? What about agricultural production? What happens to total value of agricultural production, if production falls more proportionately than the rise in prices ? Again, are the prices of agricultural commodities a measure of the prosperity of people of all occupations? Let it be supposed that they are a measure of the prosperity of agriculturists in India. But, is the whole of India agricultural, or only a part of it is so ? Evidently, the whole of India is not. Then, is what is true of a part necessarily true of the whole ? The answer is in the negative. Further, even supposing that the income of agriculturists in 1943 was much more than what it was in 1931, does that measure their prosperity ? The agriculturists are to spend also, and if in spending they pay six times more on the same items in 1943 than what they did in 1931, do they retain money with them or lose ? They do the latter. And, what is prosperity—the mere income, or the surplus income? Lastly, what is the significance of 'leaps and bounds'? 'Leaps and bounds' may be an impressive term, but is meaningless to the statistician unless he knows the bounds of 'leaps and bounds'. These

queries will suffice to understand what false generalisation may lead to, and in what direction the mind of a statistician should work to arrive at correct inference.

Let us take another example. It may be argued that the production of foodstuffs in a country in a certain year was only 0.5 per cent less than that in the previous year. There was, therefore, no *real* food shortage in the year under reference. Again, the argument creates an impression. But let us analyze it. Is the number of mouths to be fed in the country in the particular year the same as it was in the previous year, or has it increased? If it has increased, the demand for foodstuffs, other things being equal, would also increase. Again, were the foodstuffs exported from the country in the previous year? And, have they been exported in the year under consideration? If they were not exported in the previous year and have been exported in the particular year, or if the exports in the particular year exceed the exports in the previous year, the shortage of 0.5 per cent would increase to a higher figure so far as actual consumption is concerned. Further, were foodstuffs imported into the country in the previous year? If yes, have they been imported in the particular year under study? If not, there would be shortage for purposes of consumption. And, if the imports in the year under consideration are much below those in the previous year, shortage for consumption would result. These points would serve to make it clear that the task of interpretation is not strewn with roses ; it needs an analytical approach.

Another example may be found in the argument that since the quantity and value of goods imported into a certain country have been increasing, the country is prospering. Supposing the figures of imports are correct, the question arises: how much of the imported goods are being re-exported? Suppose all are being retained within the country. Then, is the consumption of goods made in the country increasing, constant or decreasing? If it is increasing or constant, the *per capita* consumption of foreign and native goods is increasing, and if increased *per capita* consumption is a measure of prosperity, the country is prospering. But *per capita* consumption would increase, remain constant or decrease according to the change that takes place in the population of the country. If the

country is being colonized and the increase in the number of immigrants far outweighs the proportionate increase in the imports, it is difficult to say that the prosperity of the country is increasing. Again, if the consumption of native goods, population remaining the same, decreases, the imports may be just sufficient to recompense this decrease, and, therefore, *per capita* consumption may not increase. There would then be hardly any increase in prosperity. But if the imports are in excess of the decrease in the consumption of native goods, the chances are that *per capita* consumption, and with it the prosperity, would increase. Again, let us look at the problem from another angle. If, along with increasing imports, the consumption of luxury goods made in the country is also increasing, it is difficult to conclude that *per capita* consumption is increasing. Luxury goods are consumed by the classes and not by the masses, but the classes may form only a small portion of the country's population. What is true of them is not necessarily true of the *whole* country. If the increment in their prosperity is nullified by the loss in the prosperity of the other section, national prosperity would not increase. These lines of thought would show the amount of care and caution required in interpreting quantitative data. They also show that arguments, though seemingly correct, may be highly fallacious, if they are not properly sifted and analyzed.

Wrong Interpretation of Index Numbers.

We have already dealt with the limitations of averages in Chapter IX,[3] and indicated there the fallacious conclusions to which a careless interpretation of averages might lead. Also, with regard to the use of both the general price and the cost of living indices we have given adequate caution in Chapter XI.[4] We may here take an example of wrong interpretation of index numbers of prices. If a general index series shows a rising tendency in a country, it may be argued that since price level, as indicated by the index series, is rising, there is inflation in the country. The argument here runs from effect to cause. This manner of arguing things is rather questionable and unsound. An effect may be the result of a multitude of causes.

[3] *Cf.* page 138.
[4] *Cf.* page 249.

To single out one cause from this multitude, without valid reasons and corroboratory evidence, is not justifiable. Moreover, index number reveals a change in the relative value of the standard of deferred payments and of goods in general—the two sides of the quantity theory equation. It is not to be inferred that a change in the level of prices is *necessarily* due to causes directly touching the value of the standard rather than to causes touching the value of the goods, which are being compared with the standard. The index number *merely* reveals the change in relationship ; the cause of the change is another question. Therefore, it is not safe to say that if an index series shows a rising tendency the cause for rise in price level is the increase in quantity of money pushed into circulation. Index number simply shows the tendency.

Wrong Interpretation of Coefficient of Correlation.

Co-efficient of correlation simply shows that two variables are related to each other. The value obtained for the co-efficient in a given case should be interpreted with great caution. Suppose a negative correlation is found to exist between area under jute and area under rice in West Bengal. The argument may run that the cultivation of jute in West Bengal is increasing at the expense of rice. This might be further interpreted to imply that the people of West Bengal want more jute for manufacturing jute cloth, gunny bags, sand bags, etc. than rice for direct consumption. This implication would prove incorrect, if, on further investigation, it was found that increase in area under jute was due, say, to war emergency requiring jute manufactures, or to relatively higher prices of jute, or to such climate changes in the State as favoured the growth of jute more than that of rice, and the cultivation of rice had gone down at the same time either because of rice plots having gone under a crop other than jute, or because of the relative fall in price of rice, or because of the rice-growers having joined the armed forces of the country. Thus, it does not necessarily mean that jute was being grown at the expense of rice. From the above line of arguments it follows that the interpretation of the value of co-efficient of correlation needs not only caution, but also a thorough knowledge of the facts that constitute other hypothesis.

Wrong Interpretation of Coefficient of Association.

While dealing with association of attributes in Chapter XVII[5] we made reference to partial association. We pointed out the reasons to which illusory association might be due, and the fallacious conclusions to which such an association might lead. One more example may be taken here.

It is observed, at a general election to the provincial legislatures, that a greater proportion of the Hindu Mahasabha candidates, who spent more money than their opponents, the Congress candidates, won the election than the Congress candidates who spent less. It is argued that the Hindu Mahasabha candidates won because of their having spent more than the Congress candidates. That is, there is a positive association between "spending more than the opponent" and "winning." The argument would be sound only if, on further investigation, it was found that these two attributes were not influenced by a third attribute. If a third attribute influenced them, the coefficient of association would work out to be a high figure even though "winning" and "spending more than the opponent" were not related. For instance, the policy, principles and programme of the Hindu Mahasabha might have generally carried the day, and "spending more" had nothing to do with its success.

General directions for Interpretation.

In all the above examples we did not doubt the accuracy of the data. We rather supposed that the data were correct, and had been properly collected and presented. But, even then we saw how wrong and fallacious conclusions could be drawn: In all these cases what appeared to be correct at first sight was not *necessarily* so when further investigations were made. Therefore, a statistical conclusion must be based on all possible investigations. In other words, statistical results should not be considered as the sole determinants of the value of given data. Statistical treatment affords only one method of judging a phenomenon. It is not the only method available. Therefore, conclusions based on statistical analysis mean nothing more than what figures imply. A statistician cannot be dogmatic about his conclusions. He cannot, and should not, assert that his figures tell that a certain result *must* be such and such.

---

[5] *Cf.* page 502.

It may be; it may not be. It will be, only when it is confirmed by other methods of studying the same phenomenon. This great limitation on the interpretation of statistics should, therefore, be always kept in view. Again, statistical laws are true only on an average, or in the long run. They study the norm, and not the abnormality. Statistics deals with the group and not the individual. These facts must not be ignored while interpreting and using statistical data.

EXERCISES

(1) What kind of mistakes are generally made in interpreting statistical data? Give examples.

(B. Com., Alld., 1936).

(2) What conclusions would you draw regarding the economic activities of the people living in the U. S. S. R. (Russia) from the study of figures given in the following table:—

(1928=100)

	1929	1930	1931	1932	1933	1934	1935
Industrial Production ..	126	164	203	231	250	300	369
Output of Investment goods	131	185	240	279	307	382	481
Output of Consumers' goods	122	147	172	190	200	230	274
Net Imports ..	92	111	116	74	37	24	25
Net Exports ..	114	128	100	71	61	52	45

(B. Com., Alld., 1939).

(3) How would you interpret the following table giving age-distribution per thousand of the population in 1911?

Age	Germany	France	England	U.S.A.	Japan	India
Under 10	234	171	209	222	244	276
10—20	203	166	190	198	198	192
20—30	164	158	173	187	154	178
30—40	139	148	152	146	138	142
40—50	105	127	115	106	101	99
50—60	76	104	80	72	77	61
60—70	51	77	51	43	57	36
70 & over	23	49	30	26	31	16

(4) The following is an extract from an office draft of an Annual Report of a large Public Library. Recapitulate the

essential statistical information in the form of a tabular statement and bring out impressively the comparison attempted in the Report.

Reading habits among borrowers vary from year to year. Topical events leave their impress on the number of borrowers and more particularly on the damage inflicted on books borrowed.

Whilst in 1938, only 15,000 books were lent out, the stress of events in 1939 attracted no fewer than 3,80,000 borrowers. These latter indented as many as 25,500 and either lost or damaged 800 books. In 1938, there were only 1,20,000 borrowers, 48,000 borrowing 4,000 books dealing with Section F (Illustrated News).

In Section E (Travel), 1,000 were lent out in 1938, but the number increased to 2,000 in 1939 while the number of borrowers increased from 2,000 to 10,000 and the losses diminished from 5 to 2 or by 60 per cent.

Section D is made up of Pamphlets. In 1938 issues amounted to 2,000 books and borrowers 25,000. The statistics only one year later were 3,000 books and 60,000 borrowers with 100 losses. In Section D, 40 books were lost in 1938 while in Section F, 680 losses among 2,80,000 borrowers were recorded in 1939. It may be noted that in this section the issues in 1939 were 14,000.

Section C (Biography) was in 1938 responsible for 1,000 books 2,000 borrowed and 2 losses ; but in 1939, although the books had increased in number by exactly one half, the number of borrowers remained exactly the same as before, and so also the number of losses.

In Section B (Science) there was no increase over the 3,000 in 1938 but 15,000 borrowers in 1938 increased by 3,000 in 1939. The losses were exactly double, 3 in 1938 and 6 in 1939.

Curiously enough in Section A (Fiction) the number of books and books lost which were respectively 4,000 and 20 in 1938, were reduced in 1939 to exactly one-half of these numbers. The number of borrowers also diminished from 28,000 to 10,000 in 1939. The total number of books in 1938 was 170 of which 100 were recorded in Section F.

(M. Com., Luck., 1942).

(5) Plot the index numbers of commodity prices in both Bombay and Calcutta (given in exercise 13, Chapter XV, page 447), against time on a graph paper. Study the graph attentively, and write a short review of the course of commodity prices in the two cities during the period May 1942 to February 1943.

(M.A., Agra, 1944).

(6) Interpret the data given below, and illustrate any two series by a suitable diagram:—

Continent or Country	Percentage of—			
	World land area	World cultivated land	World production of cereals	World population
Asia (excluding U.S.S.R.) ..	18.6	32.9	31.0	53.1
North America ..	17.3	21.2	21.5	8.2
U.S.S.R. ..	16.1	16.8	22.0	7.6
Europe (excluding U.S.S.R.) ..	3.7	16.3	16.0	17.9
Mid and South America ..	13.2	5.7	4.5	5.0
Africa ..	24.1	5.6	4.0	7.7
Oceania ..	7.0	1.5	1.0	0.5
Total ..	100.0	1.000	1.000	100.0

(M.A., Alld., 1952).

(7) Examine the validity of the conclusions in any two of the following:—

(*a*) It is observed that intelligent fathers have intelligent sons ; and intelligent grandfathers have intelligent grandsons. Therefore, intelligence is hereditary.

(*b*) Same as exercise (35), Chapter X, page 203.

(*c*) For families in general in a certain country, there is a positive correlation between the amount of money spent annually for food and that spent for clothing. It is argued that one is the cause and the other the effect in a direct causal connection.

(M. Com., Alld., 1946).

(8) How far do you agree with the conclusions drawn in the following cases?—

(*a*) Same as (*a*) of exercise 7, above.

(*b*) Two series—quantity of money in circulation and general price index—are found to possess positive correlation of a fairly high order. It is concluded that one is the cause and the other the effect in a direct causal relationship.

(*c*) It is observed that general death rates in two towns are identical. It is inferred from this that the populations of both the towns are equally healthy.

(M.A., Raj., 1951).

(9) Study the following figures very carefully:—

Season	Area under sugarcane in the U.P. in lakhs of acres.	Yield of cane in lakhs tons.	Quantity of cane crushed by factories in lakhs tons.	Quantity of sugar produced by factories in lakhs tons.	Quantity of Khandsari sugar produced in lakhs tons.	Quantity of Gur produced in lakhs tons.
1931-32	15.9	222.5	8.5	0.7	2.0	..
1932-33	17.9	261.2	16.4	1.4	2.2	..
1933-34	17.3	257.0	30.2	2.7	1.6	..
1934-35	18.4	275.8	36.9	3.2	1.2	17.2
1935-36	22.5	333.6	55.3	5.3	1.0	20.4
1936-37	25.2	380.2	63.0	6.1	0.8	24.1
1937-38	22.2	319.4	57.9	5.3	1.0	18.8
1938-39	16.5	145.4	35.1	3.2	1.0	7.3
1939-40	19.1	217.4	70.3	6.6	1.0	9.2
1940-41	25.4	286.4	52.0	5.1	1.2	15.5
1941-42	17.5	154.5	38.8	3.8	0.8	7.5

Write a short review, based on the above table, of the sugar economy of the U.P. during the period 1931 to 1942.

(M.A., Agra, 1944).

(10) (*a*) Same as (i) of exercise 6, Chapter VIII, page 88.

(*b*) The per capita national income for India for 1931-32 according to the estimates framed by Dr. V. K. R. V. Rao was Rs. 65/. The estimate for 1948-49 framed by the National Income Committee was Rs. 225/-. In 1948-49 India was, therefore, four times more prosperous than in 1931-32. Do you agree?

(*c*) The examination result of School X was 75 per cent. in a particular year. In the same year and at the same examination, only 400 out of a total of 600 students were successful in School Y. The teaching standard of the former school was decidedly better. Is it so?

(B. Com., Delhi, 1953).

CHAPTER XXII

INDIAN STATISTICS

Constitutional Position.

Article 246 of the Constitution of India deals with the subject-matter of laws made by Parliament and by the Legislatures of States. According to this Article, Parliament has exclusive power to make laws with respect to any of the matters enumerated in the Union List, the Legislature of any Part A or Part B State has similar power with respect to matters enumerated in the State List, both the Parliament and the Legislature of any Part A or Part B State have similar power with respect to matters enumerated in the Concurrent List, and Parliament has similar power with respect to any matter for any part of the territory of India not included in Part A or Part B States. The Union List includes, *inter alia.* Defence, Railways, Post and Telegraph, Currency and Foreign Exchange, Banking, Trade and Commerce with Foreign Countries, Meteorological Organisations, Census, Income-Tax, Customs and Excise Duties, and Entry 94 in this List covers "Inquiries, Surveys and Statistics for the purpose of any of the matters in this List". It is, therefore, obvious that in respect of the matters included in the Union List, statistics are to be collected by the Government of India. The State List, *inter alia,* includes Public Health, Agriculture, Livestock, Irrigation, Forests, Fisheries, State Taxes and Duties. The Concurrent List includes, *inter alia,* Vital Statistics, Economic and Social Planning, Trade Unions, Industrial and Labour Disputes, Social Insurance, Labour Welfare, Relief and Rehabilitation of Displaced Persons, Price Control, and Entry 45 in this List covers "Inquiries and Statistics for the purpose of any of the matters specified in List II (*i.e.* State List) or List III (*i.e.* Concurrent List)". Thus, in respect of statistics relating to the State List and the Concurrent List both the Parliament and the Legislatures of the States have power to make laws.

The organisation, collection and nature of statistics, available in India, have to be understood in the light of the consti-

tutional position mentioned above. While the collection of statistics on items like Defence, Railways, Banking, Currency, Foreign Trade, Population, etc., is the responsibility of the Centre, the collection of those relating to Agriculture, Livestock, Forests, Education is done by the States. In actual practice, there exists a certain amount of understanding between the Centre and the States whereby the Centre acts as the co-ordinating authority for the compilation and publication, on all-India basis, of even some of those statistics which are primarily collected by the States. This is so, for instance, with statistics relating to agriculture and education.

No General Statistics Act.

While economic statistics in India are being collected since the latter half of the 19th century, there does not exist to this day any Statistics Act empowering the Centre or the States to collect statistics in general. In the case of industrial statistics, no doubt, an Act called the Industrial Statistics Act[1] was passed in 1942, empowering the Government of India to collect certain types of statistics relating to industries through the State Governments. This Act is, however, permissive, that is to say, the decision to apply it to any particular State rests with the Government of that State. So far as agricultural statistics are concerned, they are the bye-products of revenue administration. Forest statistics, likewise, arise from forest administration. With the inception of economic controls in India during World War II, statutory provisions for the collection of certain types of statistics, no doubt, came into being; for instance, under the Sugar and Gur Control Order, 1950, sugar factories in India have to furnish periodical statements to the Government of India showing the production, stocks and despatches of sugar, failing which penal action can be taken against the defaulters. All Control Orders, however, are in their very nature temporary, designed to meet an emergency. Therefore, as soon as the emergency concerned is over and the relevant Control Order is withdrawn, the provision for obtain-

[1]This Act has since been replaced by the Collection of Statistics Act, 1953, passed by the Parliament in August, 1953. The latter Act provides for the collection of statistics relating to commerce, besides industry.

ing the statistics thereunder also ends. Besides, statistics collected under a Control Order are obtained for administering the control concerned and, as such, these statistics are also in the nature of bye-product of administration. For the collection of certain other statistics, though of a minor type, statutory provision is found to exist; for instance, under the Cotton Ginning and Pressing Factories Act, 1925, weekly statements of the actual number of bales of cotton pressed by the cotton pressing factories in a State are furnished by the factories to the State Authorities concerned. The State Authorities consolidate these statistics for their respective States and furnish them to the Government of India who, in their turn, are required to publish them, on an all-India basis, in the Gazette of India. Local Acts also exist for the registration of vital statistics. The fact, however, remains that there is no General Statistics Act in the country, providing for compulsory submission of data or empowering the Government of India or the States to call for statistics in general. To a large extent, therefore, official statistics in India continue to be the bye-products of administration.

Recent Improvements.

Many Improvements have taken place in recent years in so far as the manner of collection of data, agency responsible for such collection and the nature and scope of economic statistics are concerned.

In regard to improvements in methodology, in the case of population census, for instance, the method of one-night simultaneous enumeration had been substituted by the period system in 1941; the period system was continued in 1951 (with the further improvement that the results of complete count were verified by a sample survey. Again, in the case of statistics of crop production, the traditional method of estimating the production by multiplying the acreage under a crop with the normal yield and the condition factor has now been substituted, to a large extent at any rate, by the method of multiplying the acreage with the average yield per acre determined on the basis of random sampling crop-cutting surveys.

With the progressive improvements in the method of

collection of data and with the increasing resort to sampling, it is natural that the agency collecting the data now should be better qualified and trained than what it was some years ago. The statistician of today is no more the mere compiler of figures; he is a qualified scientist and so is his staff.

The scope of economic statistics has widened in two ways in recent years: Firstly, it has widened in respect of space. With the merger and integration of the former princely States, data in respect of them have also begun to be collected. For instance, the reporting area under forests increased from 146 million acres in 1946-47 to 170 million acres in 1949-50. Secondly, the scope of economic statistics has widened in so far as the number of items on which data are now collected is conerned. Railway statistics, statistics relating to education, statistics of industrial production, number of persons registered with the employment exchanges, and statistics of agricultural wages are some of the items on which detailed data are now available.

It is also significant to note that research in statistical technique is also proceeding apace in the country. The Indian Council of Agricultural Research, New Delhi, and the Indian Statistical Institute, Calcutta, have in recent years evolved the technique of random sampling crop-cutting surveys for the estimation of yield of crops. Attempts are also being made to evolve techniques for the collection of fisheries statistics, statistics of cost of production of crops, measurement of labour productivity, etc. Efforts are also being made to introduce statistical quality control in industry. Since statistics are the eyes and ears of administration, the administrator is now keen on having before him data collected by employing objective methods. An organisation called the National Sample Survey has been set up by the Government of India to collect, *inter alia*, statistics having a bearing on the economic and social life of the country. Statistics of rural income, expenditure, consumption, etc., for the year 1949-50 have already been collected and published by this organisation. An Agricultural Labour Enquiry has also been conducted on sampling basis and data have been collected on cost of living, wages, etc., in the rural areas. This information would come up handy in determining minimum wages payable to the agricultural workers under the Minimum Wages Act. Again, as a preliminary to determining

the extent to which the Reserve Bank of India could and should extend financial assistance to agriculture, Rural Credit Survey has been conducted by the Reserve Bank of India on random sampling basis. Thus, the use of objective methods in the collection of data is now increasing, and, in course of time, both the quality and content of Indian Statistics might become the envy of many countries of the world.

Statistical Organization in India.

In Chapter I historical development of the statistical organisation in India was traced. It will be remembered that for a long time the Indian statistical system was highly centralized, with the Department of Commercial Intelligence and Statistics (DCIS) in Calcutta serving as the chief Government agency for collecting, compiling and publishing all-India statistics obtained from the Provincial Governments and commercial and industrial bodies. Thus, it compiled and issued all agricultural statistics, statistics of foreign and coastal trade of India, statistics of inland trade, prices statistics, etc. This centralised institution enjoyed a unique existence, until the circumstances so shaped themselves as to lead to its substitution by a decentralised system. The gradual economic development of the country, the suggestions made by various committees and commissions from time to time regarding improvement of statistics, and, above all, the situation created by World War II, were the main factors responsible for the change. During the War, the centralized system proved too inadequate to meet the needs of administering an economy which was largely controlled. Need was felt of collecting several types of data, that were not being collected by DCIS. The statistical system was therefore, decentralized. At the Centre, this resulted in the establishment of statistical units in the different Ministries of the Government, each responsible for collecting the statistics relating to the subject-matter of the Ministry in which it was located. Most of the Ministries now have full-fledged statistical organizations numbering about 50 in all. The size of these organisations naturally varies with the amount of data collected and the stage of developmeant of the economy with which Ministry is concerned. Statistical organisations in the States are also of the decentralised type. In

several Part B and Part C States, they are of more recent origin than their counterparts at the Centre.

Decentralisation of the statistical system naturally called for a machinery for co-ordination, prevention of duplication, laying down of uniform standards and definitions, ensuring inter-State and international comparability. Accordingly, a Central Statistical Organisation was set up in 1951 under the Cabinet Secretariat. In the States also, the tendency towards setting up co-ordinating agencies has been noticed in recent years. In most of the Part A and Part B States, State Statistical Organisations, separate from the Departmental units, have been set up; they generally have a combined economic and statistical set-up and have also been assigned certain specific items on which they collect regular statistical data and interpret them. These State Statistical Organisations also maintain liaison with the Central Statistical Organisation, and, thus all-India co-ordination is sought to be achieved.

The existence of co-ordinating agencies does not impair the autonomy of the statistical organisations under the different Ministries at the Centre or Departments in the States so far as their day-to-day activities are concerned. The statistical organizations at the Centre continue to be fed directly, as before, by their counter parts in the States..

Decentralised organisation is perhaps inherent in the system of governance in our country. The States are autonomous; as already said above, certain subjects are assigned to them; naturally, they are the proper agency for collecting statistics relating to the subjects falling within their sphere. No single statistical organisation at the Centre can now efficiently collect statistics relating to the multifarious economic activities carried on in the different parts of the country. Economic activities are now widening and getting specialised. Specialists are, therefore, required to compile, publish and analyse the various economic statistics. No single organisation can provide experts in agricultural statistics, financial statistics, commercial and industrial statistics, and all. A statistical unit working in close liaison with the administrators and other experts in a Ministry becomes conversant with its day-to-day needs, and this, in itself, makes for improvement in the quality, content and analysis of the statistical data. The unit is enabled

to study the facts and figures in their proper perspective, analyze them and present to the Ministry a picture as it emerges from such analysis. There thus flows a continuous stream of economic advice as a matter of routine, which shapes and moulds the policy of the Government.

STATISTICAL OFFICES UNDER THE GOVERNMENT OF INDIA

It has been observed above that statistical organisations exist in different Ministries of the Government of India. While some Ministries possess small nuclear organisation, others possess several organisations, big and small. It is intended here to refer to the activities of only the major ones.

Ministry of Food & Agriculture.

Prior to 1948, agricultural statistics relating to land utilisation, area and production of crops, normal yield per acre, land revenue, harvest prices, plantations and livestock were being collected, compiled and published by the Department of Commercial Intelligence and Statistics, Calcutta; collection of market prices and their dissemination was being done by the Agricultural Marketing Adviser to the Government of India under the then Department of Agriculture; prices data relating to foodgrains were being collected by the then Department of Food; and, advisory work in the field of agricultural economics was being done in the then Department of Agriculture itself. In the beginning of 1948, all these functions were placed under one organisation set up under the Ministry of Food and Agriculture and called the Directorate of Economics and Statistics. Since that year this Directorate has been the sole organisation under the Government of India responsible for the collection, compilation and publication of all statistics relating to food and agriculture. It interprets and analyses them. It prepares notes and memoranda on agro-economic matters and tenders advice to the Ministry on agro-economic issues. The following are the major publications issued by it:—

1. *Agricultural Situation in India* (Monthly journal);
2. *Indian Agricultural Statistics* (Annual);

3. *Estimates of Area and Production of Principal Crops in India* (Annual);
4. *Indian Land Revenue Statistics* (Annual);
5. *Average Yield per Acre of Principal Crops in India* (Quinquennial);
6. *Indian Agricultural Prices Statistics* (Annual);
7. *Indian Agricultural Wages Statistics* (Annual);
8. *Indian Livestock Census* (Quinquennial);
9. *Indian Forests Statistics* (Annual);
10. *Bulletin on Food Statistics* (Annual);
11. *Commodity Statistics Series* including separate publications on foodgrains, tea, coffee, rubber, cotton, jute, tobacco, oilseeds, sugar, lac;
12. *Abstract of Agricultural Statistics;*
13. *Indian Crop Calendar;*
14. *Indian Agricultural Atlas;*
15. *Bulletin of Agricultural Prices* (Weekly);

Attached to the Ministry of Food and Agriculture is also the **Indian Council of Agricultural Research.** This council has a Statistical Branch, the functions of which are:—

(1) Advising the Departments of Agriculture, Animal Husbandry and Veterinary Services in the States as well as at the Centre, on statistical aspects of different schemes of research sponsored by the Council.

(2) Training of professional agricultural statisticians; for this purpose it runs certificate and diploma courses.

(3) Research, such as the evolution of random sampling technique for improving the collection of statistics of area and production of crops, livestok, fisheries, cost of production, etc.

In 1942-43, the Council started conducting crop-cutting experiments on random sampling basis for determining objectively the outturn of foodgrains. In 1949 50, it initiated surveys to assess the achievements of the grow-more-food campaign. In the beginning of the year 1953, both these items of work were transferred from the Council to the Directorate of National Sample Survey under the Ministry of Finance.

Ministry of Finance.

The Ministry of Finance consists of two Departments—the Department of Economic Affairs and the Department of Revenue and Expenditure. Under the Department of Economic Affairs is located the Directorate of **National Sample Survey.** This Directorate is responsible for collecting data on a national basis regarding production, consumption and other aspects of social and economic life, both in the rural and urban areas of India. A *General Report No. 1 on the First Round of the National Sample Survey, October 1950—March 1951,* was published by this organisation in January, 1953.

There is also a **National Income Unit** in this Department responsible for the preparation of official estimates of the national income of India. This unit has brought out the *Report on the National Income of India, 1948-49.*

An **Economic and Statistical Section** in this Department collects statistcs of interest to the Department. With this Section was amalgamated, in the middle of 1953, the **Economic and Statistical Co-ordination Section** of the Cabinet Secretariat, and the combined office was named the **Office of the Economic Adviser to the Ministry of Finance.**

The Department of Revenue and Expenditure maintains a statistical unit for the collection and publication of statistics relating to income tax. It also has a **Statistical and Intelligence Branch** and a **Company Law Statistics Branch.**

Besides the above, there is a Research Department in the **Reserve Bank of India** at Bombay, which comprises of four divisions. The division of Monetary Research produces the *Annual Report of the Bank, the Annual Report on Currency and Finance* and several monthly and quarterly reports. The Balance of Payments division maintains statistics relating to India's balance of payments. It recently conducted a census of foreign liabilities and assets in India. The Statistics division publishes the *Reserve Bank of India Bulletin* (monthly) which contains current financial statistics. The Rural Economics division conducted a rural credit survey in 1952-53. Some of the other important publications of the Reserve Bank of India are:—

(1) *Statement of Affairs of the Reserve Bank of India* (Weekly);

(2) *Statistical Tables relating to Banks in India* (Annual);
(3) *Statistical Statements relating to Co-operative Movement in India* (Annual).

Ministry of Commerce & Industry.

The **Department of Commercial Intelligence and Statistics** under this Ministry is easily the oldest statistical organisation in India. With the transference of certain items of its work to other offices, it has now been left with the compilation and publication of India's internal and foreign trade statistics only. The publications currently issued by this Department include:

1. *Indian Trade Journal* (Weekly);
2. *Accounts relating to the Foreign (Sea, Air and Land) Trade and Navigation of India* (Monthly);
3. *Accounts relating to the trade of India by land with foreign countries* (Monthly);
4. *Accounts relating to the coasting trade and navigation of India* (Monthly);
5. Accounts relating to the inland trade of India (Monthly);
6. *Annual Statement of Sea-borne Trade of India.*

In this Ministry is also located what has been known as the Office of the Economic Adviser to the Government of India. This Office compiles and publishes the weekly index number of wholesale prices, popularly known as the Economic Adviser's Index or the Official Index.

Attached to this Ministry is also the **Directorate of Industrial Statistics** which was shifted in 1952 from Simla to Calcutta to be merged with the **Department of Commercial Intelligence and Statistics.** One of its chief functions is to conduct an annual census of manufacturing industries. It collects data from the various industrial units in the country either on voluntary basis or under the authority of the Industrial Statistics Act of 1942. The Directorate also issues the following publications;—

(1) *Monthly Statistics of the Production of Certain Selected Industries in India.*
(2) *Monthly Statistics of Cotton Spinning and Weaving in Indian Mills.*

Ministry of Labour.

The **Labour Bureau** constitutes the chief statistical organisation in this Ministry. It collects statistics relating to wages, trade unions, workmen compensation, maternity benefits, industrial disputes, absenteeism, etc. It undertakes family budget enquiries and compiles cost of living index numbers in respect of 16 centres, the 'all-India average working-class cost of living index number' and the 'index of earnings of factory workers in India'. The Bureau issues a monthly publication—the *Indian Labour Gazette*. It has also taken over the biennial publication entitled *Large Industrial Establishments in India* from the Department of Commercial Intelligence and Statistics, Calcutta. Its annual publications include:

(1) *Labour Year Book*.
(2) *Annual Reports* on the Working of:
  (i) The Factories Act,
  (ii) The Trade Unions Act.
  (iii) The Workmen's Compensation Act.

Under the Ministry of Labour is also located the **Directorate-General of Resettlement and Employment.** The Directorate has a Statistical Section which compiles statistics relating to employment exchanges, labour training, employment in Government establishments, number of skilled and unskilled persons registered with the Employment Exchanges, etc.

There is an **Agricultural Labour Enquiry Branch** under the Ministry, which conducted an *ad hoc* survey of rural agricultural labour and is bringing out the results of the survey in a series of Reports.

A Statistical Section is also attached to the **Office of the Chief Inspector of Mines** at Dhanbad, which maintains statistics of the number of mines working, labour employed, wages, hours of work, output, accidents, etc., and brings out a monthly bulletin.

Ministry of Home Affairs.

In this Ministry is located the **Office of the Census Commissioner and Registrar General.** It is responsible for conducting the decennial census of population, the latest one having been conducted in 1951. This office has brought out several

papers containing some results of the 1951 census. To this Office has also been transferred the work relating to Vital Statistics, formerly being done in the Ministry of Health.

Ministry of Railways.

The Railway Board issues an Annual Report in two volumes: Volume I gives a narrative and Volume II, statistical information. The latter covers a wide range of operating statistics as well as the results achieved. The **Economic Adviser to the Ministry of Railways** publishes a monthly publication entitled *Monthly Railway Statistics.*

Other Ministries.

The Statistical Section of the Ministry of Education maintains and publishes, on all-India basis, statistics of educational institutions, number of schools and teachers, number of candidates passing out each year, expenditure on education by stages and types of education, etc.

Under the Ministry of Communications, the **Meteorological Department** issues weather statistics including rainfall and temperature on daily, weekly and monthly basis.

In the Ministry of Production there is an **Office of the Salt Commissioner**, which maintains statistics of production and distribution of salt in the country.

The Ministry of Transport has a Roads Organisation under it, wherein is located the main statistical branch of the Ministry. This branch maintains statistics concerning the different aspects of road development, such as road mileage, accidents, bridges, expenditure on roads, motor vehicles, and sale and consumption of petrol. It publishes the *Roads Statistics of India.*

Central Statistical Organisation.

This Organization is located under the Cabinet Secretariat. Its functions are:

(i) Advising the Ministries and other Government agencies on statistical matters;

(ii) Co-ordinating the statistical work of the different Ministries and other Government agencies with a view to eliminating and preventing unnecesary duplication;

(iii) Developing definitions and standards for improving national and international comparability and effecting improvements in the quality of information required by the Government.

This Organization provides national statistics to international bodies, such as the various organisations under the U.N.O. It is also responsible for bringing out the following publications:—

(1) *Statistical Abstract* (Annual), formerly brought out by the Office of the Economic Adviser under the Ministry of Commerce and Industry;

(2) The *Monthly Abstract of Statistics;*

(3) The *Weekly Bulletin of Statistics;*

(4) *Guide to Current Official Statistics,* formerly brought out by the Office of the Economic Adviser under the Ministry of Commerce and Industry.

STATISTICAL ORGANISATIONS IN THE STATES

It is not necessary to recall the names of statistical offices in the States if it is remembered that, like the Government of India, different departments in the States also collect, maintain and issue statistics relating to their respective subjects, either in the form of special statistical publications or in the annual reports. As already stated, **State Statistical Bureaus** have also been set up in several States to effect co-ordination in statistical matters between the different Government departments of their respective States, to provide technical guidance to these departments and to recommend improved methods for the collection of primary statistics by them. These Bureaus undertake specific inquiries and also collect and publish statistics on certain items assigned to them. Such Bureaus at present exist in all Part A States (excluding Madhya Pradesh) and in most of the Part B States.

We now proceed to a discussion of the nature and scope of certain important economic statistics in India.

POPULATION CENSUS

The census of population in India appears every ten years like a comet in the statistical firmament; it attracts much notice as it culminates, but passes away unnoticed. Every ten years, some time before the actual census takes place, a Census Act is passed. The Act, *inter alia,* binds every person in the country to furnish the census enumerator with the information he needs for recording his answers to the prescribed questionnaire, and also authorises a *bona fide* census enumerator to enter houses and interrogate people. And, every ten years a Census Commissioner is appointed to conduct the census, and Census Superintendents are appointed in each State to work under him. Each Superintendent has under him District Census Officers, Charge Superintendents, Supervisers and Enumerators. All these are appointed *ad hoc,* and trained and supplied with detailed written instructions. The actual enumeration of population is a function of the local administration. As such, it is done through a municipality in a town and through a *Tehsildar* in a rural area. The enumerators are drawn from various sources—school teachers, municipal clerks, Government officials, *Patwaris,* and many others—and what is significant is that they all work on honorary basis; people come forward and take their unpaid share in the work; the Government of India only pays for supervision and tabulation.

The object of the Census is to obtain certain first hand information in respect of every man, woman and child residing in the country at the time of the census. The items on which information is obtained are decided first. Schedules or slips are then printed and distributed among the enumerators; the scope and significance of each question is explained to them by the supervisory staff; they do the enumeration; their work is supervised and checked; the enumeration slips, together wtih an abstract statement, then move up to the State Census Superintendent; they are tabulated, and the census results are then ready for the State, and, in this way, for the whole country.

The Census of 1951.

Prior to the actual conduct of the Census, a house-list was prepared for the entire country at a period convenient to the

administration. Each house was given a number, which was firmly put at the gate of the house. For each house, known by its number, the house-list gave a distribution of the persons 'normally resident' there by sex and age. The first job of the census enumerator was to check up the particulars given in the house-list of his block by a personal visit to each house in the block.

The basis for the census was also 'normal residence'. The normal residence qualification is a corrollary to the 'Period' system of enumeration, which was adopted for the first time in India in 1941 and also followed in 1951. Under the 'Period' system, enumeration is not done at a fixed hour simultaneously throughout the country, but over a period of time. In the 1951 Census, the count began on the 9th February, 1951, and ended at sunrise on the 1st March, 1951, which is the reference date for the census. A period of 20 days was thus given to each enumerator for doing the house-to-house enumeration in his block. In each house, the enumerator was to enumerate every person whom he found there and who was also normally resident in that house. A person, who was normally resident in the house but was absent at the time the investigator visited the house, was also to be enumerated, unless such person had left the house before the 9th February, 1951, and was not expected to come back until after the 1st March, 1951. Such peron was enumerated at the place where he was first found. On the 28th February, 1951, immediately after nightfall, the enumerators visited all the buildings within their respective blocks, where houseless persons like wandering tribes, tramps, *sadhus*, beggars, etc., lived and enumerated them. The inmates of public institutions, such as, jails and hospitals, were also enumerated on the 28th February, 1951.

The investigators visited every house in their respective blocks during the first three days of March to carry out a final check. The object of this check-up was to bring the enumeration up to the reference date. This meant that every birth that took place in a house since it was last enumerated was added to the house and every death during the same period called for a deletion. Also, a visitor who had not been enumerated anywhere else during the 20-day period of enumeration, was enumerated in this final check-up.

Thus the enumerator visited each house thrice, once to check the house list, next to enumerate the resident population in February and last to give a final check-up between the 1st and the 3rd March, 1951. The enumerator handed over all the enumeration slips together with an abstract statement, in respect of his block, to his supervisor on or before the 6th March, 1951. The abstract statement gave the number of (*a*) displaced persons and (*b*) males and females. On the basis of these statements, provisional figures of population were published in April, 1951.

Soon after the enumeration was completed, the enumeration records were assembled and checked at 52 Tabulation Offices in the country, opened temporarily for the purpose. *Census of India, Paper No. 1, 1952,* published in the middle of 1952, contained the results of some of these tabulations, and subsequent *Papers* those of some others, while the Census Reports are still under preparation[1].

The 1951 Census covered the whole of India except the State of Jammu and Kashmir owing to the special conditions prevailing there, and a part of the Part B Tribal Areas of Assam. Sikkim was included in former censuses and was included in this census also. All persons who were alive at sunrise on the 1st March, 1951, and were present within the territories mentioned above (including territorial waters) were included in the count with the exception of (*a*) non-Indian nationals who were members of the staff of diplomatic and consular missions in India of foreign countries, and stationed in Delhi, Bombay, Calcutta and Madras and members of their families, and (*b*) non-Indian nationals on non-Indian war ships lying in Indian ports.

The Census Questionnaire contained 14 questions, answers to which were noted on an Enumeration Slip (which was of the size of 4″×4″). A specimen of the slip used in U.P. is given on the next page. For different types of answers in respect of most of the questions, digits or symbols were assigned; and, the enumerator was required to put down the relevant digit or symbol against each question in the enumeration slip. For

[1]Census report entitled *Census of India, 1951, Vol. I, India, Part I-A—Report* has since been published, in November, 1953. Most of the State reports have also been published.

instance, in reply to question No. 3, civil condition, if the answer were 'unmarried', the enumerator would write 1 in the Enumeration Slip, for 'married' he would write 2, for 'widowed' he would write 3 and for divorced 0. It will be seen from the Enumeration Slip (Q. No. 6) that along with the population census, a census of displaced persons was also taken, that the question about 'Race, Tribe or Caste' put in all previous censuses was dropped in 1951 as a result of Government's policy of giving official discouragement of community distinctions based on caste. Information about "special group" was collected—special group including a member of any scheduled caste, scheduled tribe, backward class or anglo-Indian community.

Enumeration Slip, 1951[1]

(Location code to be recorded here)

1. Name and relationship to head of household............

2*a*. Nationality............... *b*. Religion................

c. Special Groups....................................

3. Civil Condition............... 4. Age...............

5. Birthplace...

6. Date of arrival of D.P.......... District of origin in Pakistan..................

7. Mother-tongue........... 8. Bilingualism............

9. Dependency............. Employment.............

10. Principal Means of Livelihood......................
....
....

11. Secondary Means of Livelihood.......................

12. Literacy and Education.......................

13. Unemployment.......................... 14. Sex.

[1]Items 1 to 12 and 14 were common to all parts of India Item 13 was prescribed in each State by the State Government concerned.

The Census of India Paper No. 1, 1952 gives data on the following items:—

1. General population by Zones and States; also, females per 1,000 males and mean decennial growth rate;
2. General population by natural regions, sub-regions and divisions;
3. Livelihood Categories and Classes by Zones and States;
4. Rural and urban population by districts;
5. Population of Cities (with a population of 100,000 and over).

Below are given some of the interesting results of the Census of 1951:—

(1) The population of Indian Union (including Sikkim and 'estimates' for Jammu and Kashmir, but excluding the population of part of Part B Tribal Areas of Assam) in 1951 was 361.2 million persons (For the territory where the 1951 census was taken, 'estimates' for Jammu and Kashmir should be subtracted from this figure; the resulting total is 356.9 million persons).

(2) The comparable figures for previous censuses are as under:

General Population—Indian Union

Census year	Population (millions)	% increase (+) or decrease (−) over the previous census
1901	238.4	
1911	252.2	(+) 5.8
1921	251.5	(−) 0.4
1931	279.2	(+) 11.0
1941	318.8[1]	(+) 14.2
1951	361.2	(+) 13.3

During the last 59 years, our population has increased by 50%.

[1]According to *Census of India, 1951, Vol. I, India, Part I-A—Report*, page 122, a figure of 2 millions should be deducted from the 1941 census count 'as the estimated allowance for inflation of returns in West Bengal and Punjab'.

(3) The total world population in 1951 was estimated to be 2,400 millions. Indian Union thus contains about 1|7th of the world human force.

(4) U. P. had the largest population among the 28 States of India (*viz.*, 63.2 millions) followed, in order, by Madras (before formation of Andhra) with 57.0 millions, Bihar with 40.2 millions and Bombay with 35.96 millions.

(5) The density of population for all-India was 303 persons per square mile. Among the States, it was the highest in Delhi with 3,017 persons per square mile, followed, in order, by Travancore-Cochin with 1,015 persons, West Bengal with 806 persons, Bihar with 572 persons and U.P. with 557 persons to a square mile.

(6) The largest percentage increase in 1951 over 1941 took place in Delhi (90%), followed by 35% in Coorg, 24.6% in Tripura, 23.7% each in Mysore and Travancore Cochin, 23.2% in Bombay. Incidentally, these States, except Delhi are rice-eaters.

(7) Roughly 70% of the total population was agricultural; about 83% of the total population was rural.

(8) The total number of males exceeded that of females by about 10 millions There were 947 females per 1,000 males. In five States females predominated; in Madras there were 1,006 females per 1,000 males; in Orissa, 1,022; in Travancore-Cochin, 1,008; in Kutch 1,079; and, in Manipur 1,036. In these five States, incidentally, rice is the staple diet.

(9) The number of cities (towns with a population of 1,00,000 and over) increased from 48 in 1941 to 75 in 1951. The population of these 75 cities was 16.7 millions in 1941 and 24.1 millions in 1951, an increase of 36%.

(10) The most thickly populated city was Bombay with 2.84 million persons, followed by Calcutta with 2.55 millions and Madras with 1,42 millions (half of Bombay's). Delhi and New Delhi combined had 1.19 million persons.

(11) U.P. had the largest number of cities—16. Assam, PEPSU, Bilaspur, Coorg, Hiimachal Pradesh, Kutch, Tripura, Manipur, Vindhya Pradesh and Andaman and Nicobars had no cities.

(12) There were 51.3 millions* belonging to scheduled castes, 19.1 millions to scheduled tribes and 111,637 anglo Indians of which more than half (*viz.*, 57,074) were females.

(13) There were 7.47 millions of displaced persons (from Pakistan) in India.

(14) Indian Union (excluding Jammu and Kashmir and a part of Part B Tribal Area of Assam) had 558,089 villages and 3,018 towns. It had over 64 millions occupied houses, of which 54 million were rural. On an average, 5.4 persons occupied every rural house and 6.0 persons occupied an urban house. The census workers numbered 6 lakhs. The total expenditure incurred by the Government of India is estimated to be Rs. 1.5 crores. This gives an average of Rs. 41-12-0 per 1,000 persons enumerated.

The 1951 and previous Censuses.

Both the 1941 and 1951 censuses were conducted on the bais of house-list, normal residence and period system, that is, on *de jure* basis. Up to 1931, the censuses were taken on *de facto* basis—that is, the people were counted where they were found—and accordingly, the system followed was that of the simultaneous country-wide count on an appointed night. The latter system needed a considerable larger number of enumerators and supervisors than the 1951 and 1941 system and was, therefore, more costly. In 1951, on the basis of the house-list, a "National Register of Citizens" has been prepared for the first time and is preserved separately for every village and every block of each town in the office of the Registrar General. Both in 1941 and 1951, the enumeration was conducted straight on to the slips, which were later sorted to produce the tables; the slips bore digits and symbols to represent the answers which facilitated mechanical tabulation. Prior to 1941, copying had to be done from original lips; there was possibility of errors creeping in while copying. Mechanical tabulation was done for the first time in 1941 and was also followed in 1951. The results of the 1951 complete count were tested by a sample check, which revealed a net under-enumera-

*The figure is being revised upwards consequent on modification of figures for Hyderabad and Saurashtra.

tion error of 1.1%. Questions about caste were dropped, and those regarding displaced persons from Pakistan added in view of the special need for such data. The occupational classification was more scientific than in the past. Secrecy of the information supplied by a person in reply to the census questionnaire was ensured and census entries could not be admitted as evidence in any civil proceedings whatever, or in any criminal proceedings other than a prosecution for a census offence. To provide for between-census continuity, a permanent Census Act has been placed on the Statute Book and an Office of Census Commissioner and Registrar General has been established on a regular basis since the 1951 census. Formerly, the whole organisation was wound up when the census was over, so that the census came and went away unnoticed.

Some Shortcomings in Indian Censuses.

The shortcomings of the census reports are many. There is a marked lack of uniformity in the classification of occupations from census to census. A study of occupational structure of the country is, therefore, rendered difficult. Further, the distribution of industrial workers into employees and those working on their own account is not available. Nor are the parts of the year for which a worker is employed and the parts for which he is unemployed known. Under the circumstances it is difficult to formulate a sensible occupational plan for the country. Some of this information was, however, collected in the 1951 census. According to the *Census of India, Paper No. 3, 1953* independent workers outnumber both employers and employees amongst self-supporting persons in the non-agricultural classes. Thus, out of 32 million self supporting persons engaged in industries and services, 1 million are employers, 15 million employees and 16 million independent workers.

Indian age-returns are admittedly inaccurate mainly because of the ignorance on the part of the informants of their precise age. Besides ignorance, there are some psychological reasons too. The age-period of girls from 10 years to 15 years is defective in numbers because of the unwillingness of some people to admit having unmarried daughters, who, according to custom in the community or religious injunction, should have been married by then. For widowers and bachelors, parti-

cularly if they have a wish to remarry, there is a tendency to under-estimate their ages. Recently, married girls and particularly those who have become mothers tend to over-estimate their ages. The old people have an inclination to overstate their ages. Again, there is a marked preference for stating the age at a digit ending with zero or five. Enumerators are instructed to correct ridiculous returns of age. If they are conscientious, they do it by asking the person concerned questions about his age at the time when some well known event in the past ocurred. But with the limited ability of the enumerators it is highly doubtful that they are in all cases competent to detect the wrong and verify it. Indian custom permits only the female investigator to verify the information about *pardanashin* ladies but not many such investigators are appointed.

Returns for civil condition in 1931 exhibited an excess of married males over married females, whereas in the previous censuses the ratio was reverse. The reason for this was the promulgation of the Child Marriage Restraint Act (Sarda Act) in 1930, because of which those people who had married their male children under 18 years and female children under 14 years may have hesitated to disclose the truth for fear of prosecution.

Accuracy of census returns depends to a large extent on the capability of the enumerator and also on the general circumstances prevailing in the country at the time the census is taken. The census enumerator constitutes the front-line force, and if he is not given adequate training he cannot be expected to hit the mark with precision in all cases. Training of the enumerators is, no doubt, necessary; but, along with it supervisors too must also be picked up from among those possessing a knowledge of statistical methods and ways of conducting demographic enquiries.

Circumstances obtaining in the country at the time of census operators have a great part to play. The effect of Sarda Act has already been pointed out. And, at the time of the 1941 census, when the distribution of seats in the legislatures, local bodies and government services was based on the relative strength of persons belonging to different religions, it was not improbable that some people exaggerated the number

of members in their family. This factor caused over-estimation of the population in 1941. The 1951 Census was, on the other hand, acclaimed throughout the country as the First Census of Free India and, therefore, succeeded in securing public co-operation in quite an abundant measure. The net under-enumeration of the order of 1.1% in the 1951 Census, referred to earlier, was also disclosed by the post-enumeration sample check of 1950 Census in U.S.A.

Usefulness of Census.

The census is primarily meant for administrative purposes, but it also affords much valuable information for the economist, the sociologist and the businessman. The economist, for instance, can study, on the basis of census figures, the population trend of the country, its occupational structure, and the increment in urban population. Utilizing other relevant data along with these studies, he can trace the correlation between population-growth and food-supply, between occupational changes and the effect of granting protection to industries, and between increment in urban population and decay of rural crafts. The sociologist may study the possibilities of effecting reforms in respect of, say, ages at which people should marry, or arrangements that should be made to bring down infantile mortality. The fixation of electoral representation is made possible by a knowledge of population distribution. The questions of unemployment, defence, migration, family planning, etc., can be properly studied with the help of census data. Schemes of social insurance or rural welfare depend on census facts.

Businessmen do not always realize that the census reports contain information which is of an important nature for them. If they do, many of the problems they are confronted with can be properly attended to. India has a very large volume of internal trade, and every human being returned in the census is a *consumer*. To the businessman a knowledge of his consumers and their location is evidently of immense aid. Again, with a knowledge of the density of population of different areas an estimate of the arears where development of markets is likely can be made. The higher the density of a certain locality, the

greater is the market there. The selling cost will always be low since delivery service in a dense population is cheap. Further, knowing the number of inhabitants in a town and the quantity of goods the businessman had been usually selling there, it would be possible for him to compute the *per capita* consumption. And, if this *per capita* consumption shows a fall for no valid reasons other than lack of efficient push, the businessman can launch an intensive selling campaign to increase his sales. The class for which his goods are specially meant, say ladies, infants, military people, can be approached in a business-like manner. Besides, occupational statistics would tell the businessman whether a certain area is inhabited by the poor, whose purchasing power is low. He should then see if his goods are meant for such a class. If not, he would be wasting money over trying to gain a foothold in that particular area. He would also be able to gauge the present and future supply of labour on the basis of occupational statistics and make adjustments accordingly.

A transport agency, say, a railway, would find valuable information in the census reports. The area which is densely populated, or would be so populated if only the means of transport were improved or introduced, should receive the first attention of the transport authority. Such areas would also be good for advertising agencies to push their advertisement among the class of people to which their goods relate. Producers of staple commodities and manufacturers of industrial goods can equally benefit from the material collected at the time of census operations. If the population of a town falls, demand in that area will fall unless the demand of the existing population rises proportionately. Changes in the sex-ratio, in occupational structure, or in age composition are likely to affect the demand for goods. Similarly, life insurance companies may compare their estimates of expectation of life with those published in the census reports and see that their premium rates are properly drawn up. The legislators shall be able to study the necessity of framing legislative provisions for removing the ills from which society suffers upon a study of infant mortality, fertility rate, sex-ratio, infirmities. Numerous other uses of census figures can be thought of. This indicates the immense value of the census reports to different people.

VITAL STATISTICS

Vital statistics include statistics of births and deaths. Statistics of births and deaths in India have been collected under the Births, Deaths and Marriages Registration Act, 1886. This Act, however, provides for only voluntary registration of births and deaths. The present position regarding registration of these incidents is that some of the Part A States possess their own Births and Deaths Registration Acts, while others have provided for such registration by framing rules under the State or Municipal Acts for municipal towns and under Police or Revenue Acts for other areas. Under these rules registration by the public is compulsory in all the urban areas and in the rural areas of West Bengal and Madras; in the other rural areas, the responsibility for registrtion of births and deaths rests with the *Patwari, Chowkidar* or village head-man. As for the Parts B and C States, very few have any provision for registration of births and deaths.

In the absence of provision for compulsory registration of births and deaths in a greater part of the country as also owing to defective systems of registration, the vital statistics of India are of doubtful accuracy. They relate only to the area 'under registration'. In the past, such area covered only British India. The population under registration then included 75% of the total population of all India. After Independence, in 1949, the population under registration was also three-fourths of the total of Indian Union. Secondly, reporting is incomplete even in the registration areas; there are omissions in registration in the case of births, deaths and notifiable diseases; and, age and cause of death are not correctly reported. Birth rates and death rates computed on the basis of such registration are clearly under-estimates, but the margin of error in such rates is not known. On the other hand, a reasonably correct knowledge of such rates is very useful, and also necessary, for estimating the annual variation in population over the last census. Several committees and commissions have drawn attention to the shortcomings of Indian vital statistics, the most important of them being the Bhore Committee (1946). No marked improvement in these statistics has, however, been made yet. As for marriages, no registration is performed according to Hindu or Muslim rites.

Marriages thus remain unregistered for the bulk of the population.

For putting the vital statistics of India on an even keel, the first requisite is to make their registration compulsory under the law and the second is to set up a suitable wholetime organisation with a Cenntral Registering Authority, his counterparts in the States and a net work of staff spread throughout the country. This organisation should maintain an uptodate house list for every village, town and city in the country, giving information regarding name, date of births, etc. of the head of family as well as of the normal incidents of births, deaths, marriages in each house. The local offices of this organisation should act as local registering authorities from whom periodical and consolidated returns should move up to the State Authority through the District Authority. The particulars of births to be recorded should be name, sex of the child, date and place of birth, father's name, occupation, community, age of mother and order of birth. In the case of death, the particulars should include name, sex, date and place of death, age, martial status, community and occupation of the deceased, father's or husband's name and the cause of death. Among these particulars, sex, age of mother and order of birth in the case of births, and sex and marital status of the deceased in the case of death are very important for computing the gross and net reproduction rates. A knowledge of these rates is essential for study whether the population is constant, increasing or decreasing.

Vital statistics of India are now publised through the *Statistical Appendices to Annual Report of the Director General of Health Services*. This publication for 1948 and 1949 gives data for the rural and urban areas of each State separately on the following, among other items;

1. Births, by months;
2. Deaths, by months, age and sex;
3. Infantile mortality;
4. Deaths by Causes;
5. Maternity Death Rates;
6. Paccination statistics;
7. Sickness and mortality of prisoners in jails.

The following table gives all-India vital statistics:

All-India Vital Statistics.

All-India Rate per Mille		Year	Rural	Urban	Total
Birth	..	1946	28.4	34.0	29.4
		1949	24.9	36.1	26.4
Death	..	1949	15.1	20.5	15.8
		1946	18.6	21.4	19.0
Infantile Mortality	..	1947	141.1	168.6	145.6
		1949	119.2	139.2	122.8

The above data suffer from under-reporting. The Population Data Committee of the Government of India found in 1945 that the under-registration of deaths ranged from 35 per cent in some places to 55 per cent in others, with an average probably near 50 per cent.

The table below gives death rates by causes of mortality:—

All-India Death Rate Per Mille by Causes of Mortality.

Cause of Mortality		Death rate per mille	
		1946	1949
1. Cholera	..	0.5	0.3
2. Small pox	..	0.2	0.1
3. Plague	..	0.1	0.07
4. Fevers	..	10.8	8.6
5. Dysentery & Diarrhœa	..	0.6	0.6
6. Respiratory Diseases	..	1.4	1.3
7. Other Causes	..	5.3	5.0
Total	..	19.0	16.0

It will be seen that fevers take the heaviest toll of life in India, and that between 1946 and 1949 there was not only a decline in the over-all death rate but also in the death rates under almost every cause of mortality.

The work relating to vital statistics, being done in the

Office of the Director-General of Health Services, has recently been transferred to the Office of the Registrar General and Census Commissioner under the Ministry of Home Affairs. This is a step in the right direction. Let it be hoped that other measures of improvement will follow in due course.

AGRICULTURAL STATISTICS

The term agricultural statistics is capable of being used in several senses. In the strictly technical sense, the term refers to statistics relating to land utilisation. In common parlance, it refers to statistics of land utilisation and production of crops. In the broader sense, it includes all statistics having a bearing on the agricultural economy, such as land utilisation, production of crops, livestock, implements, fisheries, forests, prices, wages, land revenue, etc.

Area Statistics—Temporarily Settled Areas.

For a proper appraisal of the system of compilation of statistics of land utilisation and production of crops in India, the country can be broadly divided into two main regions, *viz.* (i) where the land revenue has been temporarily settled, *e.g.*, U.P., Punjab, Madras, and (ii) where it has been permanently settled, *e.g.*, Bihar, Orissa, West Bengal. The temporarily settled areas possess an elaborate village revenue agency which maintains agricultural statistics as part of land records, since the assessment of land revenue is based on the land-use and crops grown in each season. This revenue agency consists of village accountants called *Patwaris** (also called *Karnams* in the south, *Telathis* in Bombay and *Karmacharis* in Bihar), one in charge of every village or a group of small villages; Revenue Inspectors (also called *Kanungos*) to control and supervise the work of *Patwaris; Tehsil* Officers or *Taluk* Officers (including *Naib-Tehsildars* and *Tehsildars*) to control the work of the Revenue Inspectors; and, finally, District Officers (also called Collectors) for controlling the work of the entire collection of land revenue. The *Patwaris* are required to make field to field inspection of the villages under their jurisdiction and to note

* In U.P., consequent on mass resignation of *Patwaris* early in 1953, the agency has been substituted by *Lekhpals.*

the names of and the areas under the crops grown therein. The number of harvest inspections is two in Punjab and Bombay, three in U.P., one in Madhya Pradesh and twelve in Madras. That is, it varies according to crop seasons. As all fields in the temporarily settled areas have been cadastrally surveyed and mapped and allotted survey numbers, the total acreage under each of the crops for a given region is accurately known.

The method of thus collecting the area statistics can be called the method of complete enumeration. In principle, this method is perfect; but, in practice, it is difficult to claim that it is entirely free from errors. The main source of error, even where the *Patwari* is adequately trained for recording agricultural statistics, would be the slackness on his part to make a personal inspection of all fields; he might rely for his data merely on the report supplied by the inhabitants of the villages. This defect can be eliminated to a large extent by a strict enforcement, on the part of the supervising officers, of the rules laid down for checking and supervision. The *Kanungos,* in charge of revenue circles consisting of a group of about 100 villages, are expcted to satisfy themselves about the accuracy of *Patwari's* returns. They are also expected to check up the entries in the field register by actual inspection during their periodical visits to each village. The *Tehsil* and district officers are, likewise, expected to supervise the work of *Kanungos.* This supervisory staff is often burdened with other multifarious duties and hardly finds time to check up the data by an actual inspection of the fields. The Government of India have recently brought into force a scheme of Central supervision and random checking of the work of *Patwaris.*

Another reason for the *Patwaris'* inability to do the required field to field inspection appears to be the burden of several other types of work entrusted to him by the Government. Being the only, or, at any rate, the most important link between the village and the Government and being also one of the educated and most influential persons in the village, the *Patwari* is entrusted with several odd jobs like helping in the conduct of population and livestock census, preparation of electoral rolls, assisting in the Grow More-Food campaign, etc. It has also been found that the number of villages in a *Patwari's* charge is too large to be physically manageable. The

same has also been found true of the *Kanungo's* charge. The State Governments are alive to these difficulties and, depending on their resources, are reducing the charges and other workload of *Patwaris* and *Kanungos*.

In the temporarily settled areas, there are some portions which have not been cadastrally surveyed. No field to field inspection is therefore, possible in these areas. The estimates of acreage under crops for such portions are only approximate. They are not so accurate as those of the cadastrally surveyed areas.

Some error in area statistics may also arise from the different methods of recording area sown under mixed crops, such as wheat-gram mixture, and those of estimating the area of uncultivated patches in a cultivated field, or the area occupied by bunds.

Area Statistics—Permanently Settled Areas.

Land revenue has been determined on a permanent basis in the permanently settled areas. Such a system of revenue collection does not require the maintenance of agricultural statistics. Nor does it require revenue officers in the villages. The police *chowkidar* or the village headman collects the primary statistics. He is not trained in the crop-reporting work. There are also no senior officers, corresponding to Revenue Inspectors of the temporarily settled areas, to supervise his work. Till recently, the estimates of area under crops reported for the permanently settled regions have, therefore, largely been in the nature of guess. The Sub-Divisional Officer is required to ascertain through the village *chowkidar* the relation which the area under a crop in a particular season bears to the normal acreage under that crop in the settlement year. He passes on this estimate to the District Officer who modifies it in the light of his own personal experience. The District Officers, in their turn, pass on the estimates to the Director of Agriculture who puts down his own final figure on the basis of all the figures before him. Thus, no objective method is employed anywhere in the entire process, so that the area estimates are not quite accurate.

The right approach to improve the situation in the permanently settled areas would naurally be to create an

elaborate primary agency on the same lines as the one existing in the temporarily settled areas. Steps in this direction were taken in the States of Bihar and Orissa in 1944-45. The Government of Bihar established a primary agency of *Karmacharis* who have since been recording area statistics on the basis of complete enumeration. They have also appointed the necessary supervisory staff. Orissa gave up the scheme, perhaps, on financial grounds. They have, however, recently taken to the random sampling technique for area estimation. In West Bengal, a complete plot to plot survey of all lands and crops was carried out in 1944-45. Shortly afterwards, a system of collecting acreage statistics on the basis of random sampling surveys was introduced. This system now applies to almost all major crops grown in the State. The West Bengal Government have also appointed an extensive staff of agricultural assistants, one for each union in the State, to collect statistics of acreage and production. With the integration of the former princely States, many of which were under privileged tenure, reporting agencies are being set up in them also.

Yield Statistics—Traditional Method.

The total outturn of a crop is obtained as the product of the area under the crop and the average yield per acre. Average yield per acre, according to the traditional method, is obtained as the product of what is called the normal yield and the condition factor. The normal yield of a crop is required to be fixed quinquennially, by the Director of Agriculture, on the basis of crop-cutting experiments conducted by him. The normal yield has been defined as "average outturn on average soil in a year of average character" as deduced from a consideration of the information obtained or experiments made during the past quinquennium. This definition is vague and is incapable of being understood and applied in like manner by all concerned. Also, it has been found that the actual plots selected by the Directors of Agriculture for conducting the crop-cutting experiments are not typical. Personal bias has been known to affect the selection; no objective procedure is laid down for the choice of fields. The Royal Commission on Agriculture in India considered the practice of selecting by the

eye average crops for the crop-cutting experiments as statistically condemnable.

While the normal yield is fixed for a quinquennium, the condition factor varies from year to year depending on the extent and distribution of rainfall, moisture, temperature and cultural practices. The condition factor is determined by the *Patwari* or the Circle Officer as so many annas out of a standard number of annas, the standard number of annas being equivalent to the normal crop. In some States, the normal crop is represented by 12 annas; in some other States the standard is higher. The condition factor, also called the *annawari* estimate, is determined on the basis of visual observation of the actual crop in the field. This method is clearly subjective and is open to the personal bias of the observer. It is subject to an unknown margin of error. What is more, the lower officers make a guess of the condition factor of crop in their jurisdiction and their immediate superiors modify this guess in the light of ther personal experience, and the final figure is fixed by the Director of Land Records or of Agriculture, as the case may be, with the district figures before him. There is thus ample room for personal bias to enter the estimate at each successive stage.

This traditional method is open to criticism: it is subjective and the margin of error in the estimate of production resulting from it is not known. The condition factor over a series of years should average to a figure equivalent to the normal. It has, on the other hand, been observed that the condition factor over a series of years averages to a figure considerably less than the corresponding anna equivalent of the normal. This has led some people to presume that the yield estimates, as determined according to the traditional method, are under-estimates. It is, however, difficult to say with certainty *without external evidence* as to how far the estimates based on the traditional method are less than or in excess of the true figures. Such external evidence has been provided in recent years by the random sampling crop-cutting surveys. An important finding of these surveys is that the normal yields have been pitched at a far higher level than is warranted by crop conditions.

Even though the final estimates of crop production are now

being made on the basis of random sampling surveys in several States and for several crops, it is necessary to obtain accurate data regarding the condition factor and the normal yield, for the earlier crop forecasts. The normal yield should, therefore, be computed as a moving average of the actual average yields per acre obtained by the random sampling method during the preceding 10 years. The condition factor for a district should be computed as the weighted average of circle averages with weights proportional to areas under the crop in each circle. The State Governments are now proceeding on these lines.

YIELD STATISTICS—Random Sampling Method.

The random sampling method for the estimation of yield statistics offers itself as the most scientific procedure to replace the traditional method. Attempts in this direction were made in Bihar and Orissa as early as 1923; but nothing much came out of them. During World War II when problems arising out of production and distribution of foodgrains emphasised the need for accurate yield statistics of principal food crops, rapid measures were taken to popularise the method of random sample crop cutting surveys on foodgrains on an all-India basis. This work was done under the auspices of the Indian Council of Agricultural Research (ICAR).

The plan adopted for the ICAR surveys is one of stratified multi-stage random sampling. The plan consists in selecting at random a certain number of villages from each *tehsil* (*taluka*) of a State. The number of villages selected in each *tehsil* is approximately in proportion to the area under the crop. In each selected village, two or three fields are selected at random out of all the fields sown with the crop. In each such field a plot, usually of the size of 1/80th of an acre ($33' \times 16\frac{1}{2}'$) is located at random. The selected plot is demarcated by ropes or chains. The crop-cutting experiment proper starts at the time of harvesting. The crop is cut from the plot, threshed, winnowed and weighed. Allowance for the moisture content is made on the basis of special experiments conducted for the purpose.

The technique evolved by the ICAR has been found to be practicable in the hands of the existing departmental staff in the States. The whole field work is now carried out by the

staff of the Revenue or Agriculture Department under the supervision of technically trained officers attached to the Department; the ICAR provide the technical guidance wherever necessary.

It will be seen that the random sampling method is a considerable improvement over the traditional method of crop estimation. The need for determining the normal yield quinquennially and the condition factor annually has been done away with; the random sampling method gives the average yield per acre directly. The method is not only objective from the start to the finish, but also possesses the merit of providing the margin of error in the estimate. The principle of random sampling ensures that the sample selected out of the total population of all fields under the crop would reflect all the characteristics of the population. That is to say, factors like irrigation, types of varieties of soils, different cultural practices, etc., are represented in the sample in their due proportions.

The first series of crop-cutting surveys under the above system were initiated by the ICAR in the *Rabi* season of 1943-44 on wheat in Punjab and U.P. and in the *Kharif* season of 1944-45 on paddy in Orissa and in one district in each of the States of Madras, Bombay and Madhya Pradesh. It has since been extended to almost all foodgrains in all States where they are important. In 1951-52 , statistics of food production were available for 70% of the area under food crops in the country by this scientific method. The method is also being applied to crops other than foodgrains, for instance, to cotton, jute and certain oilseeds. The method has also been officially adopted by the Government inasmuch as official estimates of final production of crops are now based on this method for such crops and in such States where the experimental stage of crop-cutting surveys by this method is now over.

In West Bengal, the Indian Statistical Institute started random sampling surveys in 1943-44 not only for the estimation of yield but also for that of acreage, West Bengal being a permanently settled area. The State Government have now officially adopted the random sampling technique. The technique is somewhat different from that of the ICAR, particularly in respect of the shape and size of the experimental plot.

A comparison of the estimates of production based on the traditional method with that based on the random sampling method would show wide divergence in certain years in some States. The following table for rice provides an illustration:

Comparison between ICAR and official estimates of production of rice.

State	Percentage over-(+) or under-(−) estimation of official estimate over ICAR estimate.			
	1946-47	1947-48	1948-49	1949-50
Assam ..	..	− 1·6	− 0·3	+ 7·6
Bihar ..	−30·7	−34·5	−26·9	+ 3·0
Bombay ..	− 4·5	+ 4·8	− 7·2	+ 7·6
Coorg ..	− 9·1	−14·2	−18·1	−20·0
Madhya Pradesh	−20·0	− 7·9	− 9·5	−11·3
Madras ..	+12·6	− 4·1	− 0·2	+ 3·8
Orissa ..	−11·1	−18·2	−20·3	−13·7
U. P. ..	+ 7·3	+ 0·9	+ 2·8	+27·0
Total	− 8·0	−12·7+	− 9·5	+ 2·7

On the whole it will be seen that the estimates based on the traditional method are usually under-estimates. The overall divergence between the two sets of estimates has, however, been narrower in recent years.

Land Utilization Statistics.

These statistics are published for the States in Volume I of *Indian Agricultural Statistics* and for the districts in Volume II of the same publication. Prior to Partition, Volume I used to give districtwise information in respect of the then Provinces and Volume II in respect of the Indian States.

Land utilisation statistics are based on the acreage returns prepared by the village acountant for revenue purposes, except in the permanently settled areas where, as already pointed out above, no village papers exist. These statistics are not yet available for the entire country. They are not collected for areas which are inaccessible (as, for example, the hilly tracts

of Assam, Manipur and Himachal Pradesh) or areas where no village agency exists as the *jagir* areas and areas held under privileged types of tenures in the former Indian States (as, for example, parts of Rajasthan, Saurashtra and Madhya Bharat) and for the enemy-held areas of Jammu and Kashmir. Thus, in 1949-50, land utilisation statistics were available for 679 million acres out of a total geographical area of 811 million acres. That is, about 16% of the total geographical area of the country was non-reporting for these statistics.

Broadly speaking, land utilisation statistics are available under the following heads:—

(1) Classification of area.

(2) Area sown, crop-wise.

(3) Irrigated area.

Statistics of classification of area are available under five heads. These heads and the proportion of the 679 million acres of reporting area falling under each of these classes is given below:—

	Percent. of Reporting Area.
(1) Area under forests	15
(2) Area not available for cultivation	17
(3) Other uncultivated lands excluding current fallows	17
(4) Area under current fallows	9
(5) Net area sown	42
Total	100

Thus, of the total reporting area, about 51% was the cultivated area in 1949-50, 42% being sown to crops and 9% lying fallow to recoup its fertility.

The above classification affords only a broad outline of how land is utilised in the country. It does not give area under minute categories of land use, which is very necessary for agricultural planning. With effect from 1950-51, the Government of India have, therefore, increased the number of land-use

classes from 5 to 9. The existing classification is compared with the revised classification below:—

Existing Classification.	Revised Classification.
1. Forest.	1. Forest.
2. Not available for cultivation.	2. Land put to non-agricultural uses, 3. Barren and uncultivated land.
3. Other uncultivated land excluding current follows.	4. Permanent pastures and other grazing land, 5. Miscellaneous tree crops and groves, 6. Culturable waste.
4. Current follows.	7. Other fallows, 8. Current fallows.
5. Net area sown.	9. Net area sown.

In either classification, the sum of the net area sown and the area sown more than once would be equal to the total cropped area, that is, the total area sown under different crops. In 1949-50, the gross sown area was 321 million acres.

Statistics of area under crops are fairly detailed. They are available individually for almost all important crops. The crops are divided into two main groups: food crops and non-food crops. The first group includes cereals, pulses, sugarcane condiments and spices and fruits and vegetables. The second includes oilseeds, fibres, dyes and tanning materials, drugs and narcotics, fodder crops, green manure crops and others. In 1949-50, food crops occupied 82% of the 321 million acres of total cropped area, only 18% being devoted to non-food crops.

Statistics of area irrigated are available both according to the sources of irrigation and according to crops irrigated. The former gives the net area irrigated and the latter the gross area irrigated. If two crops are irrigated from the same source of irrigation, say a canal, on an acre of land in the same year, the net area irrigated is 1 acre while the gross area irrigated is 2 acres. The sources of irrigation separately enumerated are:

(1) Government canals.
(2) Private canals.
(3) Tanks.
(4) Wells.
(5) Other sources.

The net irrigated area in 1949-50 was 50 million acres, of which 20 million acres were irrigated by canals and 14 million acres by wells. The gross irrigated area was 54 million acres, of which 24 million acres were under rice alone, 2·5 million acres each under sugarcane and fruits and vegetables and 1 million acres under cotton. Viewed from another angle, 32% of the total area sown to rice was irrigated; this percentage was 67% in the case of sugarcane and 8% in that of cotton. That irrigation facilities are meagre in the country can be judged from the fact that only 16·8% of the total cropped area irrigated in the country in 1949-50.

It should be remembered that the area under a crop in India represents the area actually sown irrespective of the fact whether the crop reached maturity or not, except in cases where after the failure of the first crop, the land was devoted to another crop, when the area would be recorded under the other crop sown. In some foreign countries area under crops stands for the area harvested. In some countries, statistics are available of both the area sown and the area harvested.

A possible source of inaccuracy in area statistics in India arises from the varying practices followed in different States for recording the area under crops sown mixed. The present method of estimating such area consists in recording for each field the area covered by the mixed crop, totalling such area for the district and allocating the total to the component crops in accordance with the fixed ratios prescribed by the State authorities. These ratios are fixed only for principal mixtures and, in the case of minor mixtures, the area under the mixture is entirely assigned to the area under the principal crop. Another method in vogue is that the *Patwari,* in the course of his field inspection, makes a judgement by the eye of the ratios in which the mixtures are grown and apportions the gross area under the mixture to the component crops. Both these methods are attended with some inaccuracy. The Government of India have taken steps to remedy their defects.

Another defect in area statistics might arise from the fact that a particular region has not been surveyed and mapped, or that the data in respect of a region is not based on complete enumeration but is estimated, or that an *ad hoc* estimate is made for an area which neither has a reporting agency nor has been surveyed.

Area and Production of Principal Crops.

The Government of India now issue crop estimates for 23 commodities which may be grouped as under.

Cereals: Rice, jowar, bajra, maize, ragi, wheat and barley.

Pulses: Gram, pulses (tur, other *Kharif* pulses and other *Rabi* pulses).

Oilseeds: Groundnut, sesamum, rape and mustard, linseed and castor seed.

Fibres: Cotton, jute and mesta.

Miscellaneous: Sugarcane, potatoes, tobacco, pepper, ginger and chillies.

Besides the above crops, statistics of area and production are also available for plantation crops (which include tea, coffee and rubber) and for certain minor crops such as small millets, cocoanut, sann hemp, indigo, opium, lac, tumeric and subsidiary foods (which include papaya, sweet potatoes and tapioca).

The above statistics of area and production of crops are available in the *Estimates of Area and Production of Principal Crops in India* issued by the Directorate of Economics and Statistics. This publication also gives the average yield per acre of the crops and rainfall data.

The area statistics in this publication have generally the same reliability as those in the *Indian Agricultural Statistics*. In some cases it may be found that the area figures under a particular crop given in the two publications are not identical. This may be due to one or more of the following reasons:—

(*a*) Time-lag between the preparation of the two sets of figures;

(*b*) Inclusion of certain area in one set of figures and exclusion of the same from the other set. This may particularly

happen when a crop is commercially important and, therefore, even *ad hoc* estimates for it may be included in the crop estimate.

(*c*) Difference in the methods of recording area under mixed crops. In U.P., for example, area under oilseeds sown mixed with pulses is assessed differently for the crop estimate from the *Indian Agricultural Statistics*.

Area within a State for which the State Government does not prepare regular crop estimate is called Non-Reporting Area for that crop estimate. Likewise, a State which does not prepare estimate for a particular crop for which all-India crop estimate is issued is also called Non-Reporting Area. It can be imagined that prior to integration of Indian States, many of them must have been Non-Reporting Areas. With the extension of reporting agency in the Indian States merged with former Provinces and also with the creation of Unions of States (called Part B States), the Non-Reporting Areas gradually thinned out and became almost negligible by 1952-53. It should also be clear that as a result of (*a*), the gradual extension in the Reporting Area, (*b*) changes in the method of estimation of area and production in West Bengal and Bihar and (*c*) change-over from the traditional method of estimation of yield to the random sampling technique in respect of certain crops in several States, the area and production figures for the recent years, say since 1949-50, are not strictly comparable with those for the previous years.

Crop Estimates.

All-India Crop Estimates formerly called Crop Forecasts, are at present issued for 23 crops, already mentioned in the foregoing section. They are issued in the form of bulletins, one for each crop; they are published in the daily newspapers and economic journals and are broadcast over the radio.

Generally three estimates are issued for each crop. In respect of castor seed, pepper and ginger, however, only one estimate is issued in a year, for, *jowar, bajra, maize, ragi,* jute, 'other *Kharif pulses*' and 'other *Rabi* pulses' two each, and for wheat and cotton five each.

The First Estimate is generally issued about a month after the completion of sowings. It is intended to give an idea of the area sown to the crop and of how germination is proceeding. The Second Estimate genrally follows about a couple of months later and gives information about the extent of area sown (including the late sown area), the condition of the crop and also the expected yield in the case of some crops. The Final Estimate (Third Estimate where only three Estimates are issued, Fifth Estimate where as many as five Estimates are issued) contains data about the total area sown and the yield harvested or expected to be harvested.

For the estimates issued prior to the Final Estimate, the primary reporting agency, that is, the *Patwari* is required to furnish only rough estimates of the area sown to the current crop as compared with the crop of the previous year. For the Final Estimate in the temporarily settled areas, he has to undertake a field-to-field investigation before furnishing the return. According to the traditional method, discussed earlier, the *Patwari* or the *Kanungo* also makes an *annawari* estimate of the size of the crop at the Final Estimate stage and furnishes the same to *tehsil* authorities, from where it goes to the Director of Agriculture *via* the District authorities. Where random sampling crop-cutting surveys are conducted to estimate the average yield per acre, the Director of Agriculture gets the yield figures from his statistician.

The Final Estimates are really not final; they are subject to revision, if found necessary, in the light of subsequent information. Such revision, however, is done at the time of issuing the Final Estimate of the next year's crop, when the Final Estimates of the previous year are given the name of Revised Estimates.

The table below compares the estimates of area and production of some of the principal crops in India in 1952-53 with 1951-52:

Area and Production of Principal Crops—India.

Crop	Area ('000 acres) 1952-53	Area ('000 acres) 1951-52	Production Unit of Weight	Production 1952-53	Production 1951-52
Rice ..	74,674	73,665	'000 tons	23,424	20,741
Wheat ..	24,041	23,450	—Do—	6,762	6,039
Other Cereals	1,01,081	95,124	—Do—	17,398	15,660
TOTAL CEREALS ..	1,99,796	1,92,239	—Do—	47,584	42,440
Gram ..	17,267	16,857	—Do—	3,771	3,293
Other Pulses ..	29,402	29,522	—Do—	4,715	4,956
TOTAL Pulses ..	46,669	46,379	—Do—	8,486	8,249
Sugar ..	4,376	4,792	—Do—	5,260	6,068
Cotton ..	15,678	16,198	'000 bales	3,050	3,133
Jute ..	1,834	1,951	'000 bales	4,695	4,678
Oilseeds ..	27,511	28,388	'000 tons	4,635	4,820

Production of sugarcane is in terms of *gur* or raw sugar: that of groundnuts included in oilseeds is in terms of nuts in shell. A bale of cotton weighs 392 lbs. and that for jute 400 lbs. The figures for 1952-53, being Final Estimate figures, are subject to replacement by Revised Estimates when Final Estimates for 1953-54 will become available.

The above table hows that while the area and production of cereals and pulses in 1952-53 have gone up as compared with 1951-52, those under the commercial crops have declined except in the case of the production of jute which shows a nominal increase over the previous year.

All the short-comings and limitations of area and yield statistics, mentioned in the foregoing pages, apply with equal force to Crop Estimates. An added shortcoming is that these Estimates hardly ever issue on the dates prescribed for them. In several cases they become available only after the crop, at any rate the major portion of it, has been marketed. Delays in the issue of all-India estimates become unavoidable, if the preparation of the Estimate in any State, particularly a major State, is held up beyond the due date. The ultimate cause for delay is, however, traceable to delays at the primary stage of

reporting and for this one remedy is to reduce the workload on the *Patwari*. Delay in the preparation of State Estimates also results from the adoption of the random sampling method of yield estimation: whereas in the traditional method the condition factor (*annawari estimate*) was determined by the eye when the crop was standing, ready for harvest, the average yield per acre according to the random sampling technique can be estimated only after the crop in the experimental plot has been but, threshed, and winnowed and weighed. These delays explain why instead of being called Crop Forecasts, these figures have now been designed Crop Estimates.

Shortcomings and Improvements in Agricultural Statistics.

The defects in agricultural statistics and improvements effected and proposed have already been mentioned in the foregoing pages. They may be summarised here. The chief shortcomings of Indian agricultural statistics may be classified as under.—

(*i*) Gaps in coverage;
(*ii*) Lack of uniformity in definition and classification;
(*iii*) Defects of tabulation and processing;
(*iv*) Defects of the primary reporting agency;
(*v*) Defects of inspection, supervision and checking;
(*vi*) Defects of planning and coordination; and
(*vii*) Delay in publication.

Gaps in coverage.—As already pointed out, agricultural statistics are not, at present (1949-50), available for 132 million acres of the geographical area of India. Leaving out an area of about 54 million acres on account of the special circumstances obtaining in the State of Jammu and Kashmir, the balance of 78 million acres needs to be brought on reporting basis. Of this 78 million acres, Rajasthan has 28 million acres, Assam 21 million acres and Saurashtra 8 million acres.

Most of the non-reporting areas, as well as some of the reporting areas, are not cadastrally surveyed and this constitutes a chief impediment in the way of proper organisation of agricultural statistics in these areas. Further, there are certain areas which, though surveyed, are non-reporting for want of any reporting agency existing there. Estimates of production for a number of crops, though minor ones, are also not

available. Fruits and vegetables, condiments and spices, fodder and cattle feeds are cases in point. The agricultural statistics of India can be said to be complete only when estimates of acreage under different land-use classes and under different crops become available in respect of the entire geographical area of the country, and estimates of production become available for every acre of land cropped.

It is significant to note that much has been done during the last few years to remedy the above defects. The reporting area has increased from 566 million acres in 1946-47 to 679 million acres in 1949-50 and is likely to increase in future till the whole country is covered; reporting agencies have been set up in many areas which did not have them; and, regular estimates are being published for several minor crops for which they were not available prior to, say, Partition. For cereals, the production estimates include even the non-reporting areas now. The position is, therefore, one of continuing improvement.

Lack of Uniformity in Definition, etc.—The method of obtaining area statistics of temporarily settled areas being different from that in vogue in the permanently settled area, the acreage statistics of the country are not of uniform quality. The position in this regard has, however, improved considerably in Bihar and West Bengal, and attempts are being made to improve the position in Orissa.

The definitions of the different land-use classes were governed generally by local custom and usage and were, therefore, not uniform. or example, the period for which lands lying fallow were included under current fallows varied from 2 years in Punjab to 10 years in Bombay. Along with increasing the number of land-use classes from 5 to 9, the Government of India have now laid down uniform definitions for all States, and therefore the future land-utilization statistics will be available on a uniform basis.

The methods of yield estimation are not uniform. While in some States, the *annawari* system is in vogue, in Punjab the method of direct estimation is followed whereby the yield is determined directly in terms of maunds per acre; in the case of some crops the random sampling technique is being adopted in most of the States. This, however, is bound to be the

position in a period of transition from the traditional method to that of random sampling. The methods of estimation of normal yield and condition factor also need to be improved, as mentioned earlier.

Some divergence in the methods adopted exists in respect of recording areas under mixed crops, of uncultivated patches in a cultivated field and of area under bunds. Certain schemes for remedying these defects are, however, under way.

Defects of Tabulation and Processing.—A good deal of data now collected runs to waste for want of proper processing. For instance, in Punjab, Delhi and Madhya Pradesh information regarding transfers of agricultural property is collected for each village by the *Patwaris;* but, it is not consolidated beyond the *tehsil* stage and, hence, no figures for the State as a whole are readily available. In other States, though similar information could be abstracted from the existing records, this is not done. Again, while data regarding ownership holdings is compiled and tabulated once in five years, no attempt is made to compile similar tables for cultivators' holdings.

Defects of the Primary Reporting Agency.—These have already been pointed out, both in respect of the temporarily settled and permanently settled areas. Even though area statistics in the temporarily settled areas are more reliable, there is need for checking and verification of *Patwari* records. It is not unlikely that though the *Patwari* may record the total area correctly, his record of the break-up by crops may be different from the actual. One of the chief causes for the inaccuracies in the reporting of the primary agency is the large increase in the work-load of the *Patwari*. State Governments are now taking steps to reduce the charge of the *Patwaris*. Neglect and bias of the primary reporting agency also result in inaccurate figures. During the days of food controls, which have now been considerably relaxed, wherever the *Patwari* was responsible for procurement it was not unlikely that he underestimated production only to soften the rigours of procurement.

Lack of Supervision, etc.—The *Kanungoes* and *Tehsildars,* who are touring officers, are expected to visit a certain number of villages and inspect the work of *Patwaris*. In the midst

of their heavy administrative duties they are not always able to devote personal attention to such inspection. This apathy on their part induces the *Patwari* to be negligent. Also, instead of paying a surprise visit, the supervising officers intimate their tour programmes well in advance to the *Patwaris* so that the very object of inspection is defeated. The State Governments are now insisting on better supervision and checking of the work of primary agency, and villages to be visited are even being selected on random sampling basis. They are also reducing the charge of *Kanungos* to make their jurisdiction manageable. Schemes of Central supervision and checking are also in force now.

Lack of Coordination, etc.—In several States the Departments of Agriculture and Food or Civil Supplies obtain estimates of food production independently of each other. The former is concerned with the Grow-More-Food activities and the latter with procurement and equitable distribution of food: often, therefore, the estimates of the two Departments run at variance with each other.

Delay in publication.—This is a common defect with a large majority of official statistics. In the case of agricultural statistics delay appears to be inherent in the very system of collecting them: The data are first recorded for each survey number in *Patwari's* books; they are then consolidated, first for each village, then for each *Tehsil* at the *Tehsil* headquarters, thereafter for each district at the district headquarters, then for the State as a whole at the State headquarters; and each State sends its consolidated return to the Directorate of Economics and Statistics at the centre for consolidation and publication on all-India basis. In this chain, if the returns for one village or a *Patwari* circle in a *Tehsil* are delayed, the return for the entire *Tehsil* will be held up, which in turn will hold up the all-India publication. An alternative that suggests itself is stricter insistence on punctuality at every stage of the hierarchy, failing which the only way out would be the consolidation of even the primary returns at the Centre.

Season and Crop Reports.

These *Reports* are issued by all the Part A States and Mysore at present. These *Reports* give, on districtwise basis,

the land utilisation statistics, statistics of area and production of crops, livestock statistics and data on agricultural prices, etc., wih a narrative on the weather and crop conditions. The information given in these *Reports* on different types of statistics should generally be found to tally with the all-India publications mentioned above, *viz.*, *Indian Agricultural Statistics and Estimates of Area and Production of Principal Crops in India.* In some cases, however, exact tallying may not be found to occur, because the *Season and Crop Report* may be based on revised information received from certain districts subsequent to the State having sent the available information to the Central Government for publication in the all-India publications mentioned above.

Livestock Statistics.

Livestock statistics are collected through the quinquennial census of livestock, poultry, agricultural machinery and implements, etc., conducted in the States. The last census was conducted in 1951. An idea of the nature of information in respect of livestock and poultry collected through the census of 1951 may be had from the following:—

1. BOVINES	(In thousands)
(1) Bulls over three years	
(*a*) cattle	61,759
(*b*) **buffaloes**	6,778
(2) Cows over three years	
(*a*) cattle	49,848
(*b*) buffaloes	21,842
(3) Young stock	
(*a*) cattle	43,492
(*b*) **buffaloes**	14,731
TOTAL	
(*a*) cattle	155,099
(*b*) **buffaloes**	43,351

II. OTHER LIVESTOCK

(1) sheep	..	38,829
(2) goats	..	47,077
(3) horses and ponies	..	1,514
(4) donkeys	..	1,239
(5) mules	..	60
(6) camels	..	629
(7) pigs	..	4,420
	TOTAL ..	93,768

TOTAL LIVESTOCK

(a) Cattle	..	155,099
(b) Buffaloes	..	43,351
(c) Other Livestock	..	93,768
	TOTAL ..	292,218

III. POULTRY

(1) Fowls	..	67,135
(2) Ducks	..	6,264
	TOTAL ..	73,399

The data collected through the livestock census are published in the *Indian Livestock Census*. Besides containing data on population of livestock and poultry, this publication also contains information on the number of agricultural implements and machinery of different types, including tractors, in the various States. The data available in the 1951 Census cannot be said to be strictly comparable with those for the previous years, inasmuch as the previous censuses related to India before Partition and integration of States. Also, in the censuses prior to 1951, some of the then existing Provinces and Indian States did not participate. The agency for conducting the census is again the village revenue staff in the temporarily settled areas, and other local agencies (*e.g.*, veterinary staff)

in the permanently settled areas. In urban areas the enumeration is done by the local bodies. Since the conduct of this Census is hardly preceded by intensive training of the primary reporting agency, it is difficult to vouchsafe the accuracy of the data collected.

Some States undertake annual enumeration of livestock. These annual data, together with the quinquennial data for those States which conduct only the quinquennial census, as also the statistics of ploughs, carts (which until 1946-47 were published in the Indian Agricultural Statistics) are published in the *Indian Livestock Statistics*. This publication also gives statistics of output of livestock products such as milk, butter, *ghee*, meat, eggs, hides and skins and wool. It also gives data on foreign trade in livestock and livestock products, utilisation of livestock products, etc.

Forest Statistics.

Upto 1946-47, statistics relating to forestry were published in *Annual Returns of Statistics relating to Forest Administration in India*. Since 1947-48 the name of the publication has been changed to *Indian Forest Statistics*.

The forest statistics covered in this publication relate to area under forests, volume of standing timber, outturn of timber and other forest produce, employment in forestry and forest industries, revenue and expenditure of Forest Departments in the States, and foreign trade in forest produce. The data relate to financial year ending with 31st March. Prior to Independence, these statistics were available for the former Provinces and some Centrally Administered Areas. Now they are available for all the States of the Indian Union.

A *Review of Forest Administration* in the country is also prepared on quinquennial basis and is published along with the *Indian Forest Statistics* every five years. The *Review* throws light on the fluctuation in the area under the control of Forest Departments, progress of forest research, and financial results.

It will be recalled that one of the classes under which land utilisation statistics are published is 'forests'. The area under forests given in the *Indian Agricultural Statistics* does not in all cases tally with the area under forests given in the *Indian*

Forest Statistics. The main reason for this discrepancy is that the term 'area under forests' is not used in the same sense in the two publications. In the *Indian Forest Statistics* the whole area, including even the non-wooded area, under the control of the Forest Department of a State is taken as area under forests in that State. In the *Indian Agricultural Statistics*, on the other hand, the cultivated or fallow area within a forest area is included in the appropriate class. In other words, while the *Indian Forest Statistics* serves an administrative need and indicates the ownership of forest areas, the *Indian Agricultural Statistics* is based on the actual utilization of land. It may also be added that while the former relates to the financial year (April to March), the latter relates to the agricultural year (July to June).

FOOD STATISTICS

Controls on the sale, movement, prices and consumption of foodgrains were imposed in India during World War II when acute scarcity of food was felt in the country owing, *inter alia*, to cessation of rice imports from Burma. These controls were first imposed in 1943-44 and although they were intended to tide over the temporary emergency created by the War, they were continued even after the War because of damage to crops in the country and disturbed economic and political conditions. Since the latter half of 1952, however, they have been progressively relaxed.

Since the imposition of food controls, a large volume of statistical data has come to be collected by the States and the Central Government, primarily to help in their administration. These data are contained in the *Indian Food Statistics*, 1949, and in the annual *Bulletin on Food Statistics* brought out since 1949 by the Directorate of Economics and Statistics of the Ministry of Food and Agriculture. Broadly, the data given in these publications cover production of foodgrains; procurement of foodgrains by the State Governments for their own consumption as well as, of course in the case of surplus States, for supply to deficit States; imports of foodgrains from abroad; stocks of foodgrains with the Central and State Governments; despatches of foodgrains from surplus States and

Central Government's stocks, and receipts by the deficit States under the Basic Plan; number of persons rationed or otherwise dependent on Government distribution; size and composition of rations; offtake by the rationed population; controlled prices; financial aspects of food distribution including consumer subsidies.

Food Balance-Sheet.

The following table gives the all-India balance-sheet of the availabilities of foodgrains with the Central and State Governments and their distribution for consumption:

Food balance-sheet.

(Thousand Tons).

	1949	1950	1951	1952
1. Opening stock ..	1,044	1,581	729	1,307
2. Procurement by States	4,610	4,617	3,766	3,416
3. Supplies to States out of imports ..	3,619	2,085	4,679	3,474
4. Offtake from Government stocks ..	7,692	7,554	7,867	6,696
5. Closing stock ..	1,581	729	1,307	1,501*

It will be seen from the above table that with the initiation of relaxations in food controls in the middle of 1952, the offtake from Government stocks went down by about a million tons during the year 1952 as compared with any of the three previous years, and the stocks with the Central and State Governments taken together, *viz.*, 1.9 million tons, were the highest ever.

Imports of Foodgrains.

The table on next page gives the imports of foodgrains on Government account and their value during the last five years:—

* This quantity was held by the State Governments. The Central Government held a reserve of 410,000 tons.

Quantity and Value of Imports of Foodgrains.

Year		Quantity ('000 tons)	Value (Crores of Rs.)
1948	..	2,841	129.7
1949	..	3,706	144.6
1950	..	2,125	80.6
1951	..	4,725	216.4
1952	..	3,864	209.7

It will be seen that the imports in 1951 were more than double of those in the previous year. These large imports in 1951 were necessitated by the serious damage to cereals crops caused in 1950-51 owing to devastating floods followed by prolonged drought between July-December, 1950. The burden of food imports on the Indian exchequer can be very easily seen from the column giving the value of imports in the above table.

INDUSTRIAL STATISTICS

India has long been an agricultural country; she retains her essentially agricultural character even to-day. For some time, however, it has been widely realised that a rapid programme of industrialisation is an indispensable requisite for her balanced economic development. The Five Year Plan, and the several plans that preceded it, bear testimony to this. Industrial statistics occupy an important place in such a programme.

The growth of our industries being recent, statistics of industrial output are not available over a sufficiently long period of time except for cotton and jute. Even for the few large-scale industries that were developed after World War I, no coordinated attempt was made to collect output statistics till the early thirties. Prior to the outbreak of World War II such statistics were collected for about a dozen large-scale industries; and, except for cotton manufacturers, sugar and other excisable commodities they were based on returns supplied voluntarily by the industries concerned. Thus the available statistics pertaining to industries were incomplete and of doubtful accuracy, and were inadequate for purposes of economic planning.

Annual Census of Manufactures.

In 1942, the Government of India enacted the Industrial Statistics Act, which gave them statutory powers to collect, *inter alia,* statistics relating to industries which are within the purview of sections 2(*j*) and 5(*l*) of the Factories Act, 1934, and to conduct statutory census of manufacturing industry. The term 'Factory', as amended by the Factories Act, 1948, includes all concerns which either employ 20 or more persons, or employ 10 or more persons and use power. The Act provides for the collection of returns in prescribed forms from occupiers of factories, failure to furnish which is punishable by fine. Though the Act was passed in 1942, it was only late in 1945 that the Directorate of Industrial Statistics was set up at the Centre to enforce it. The first step towards implementing the Act was for the State Governments, through whom the census is conducted, to notify the rules for conducting it. This was done in 1945, and the first census was conducted in 1946. Annual censuses have been conducted since, and the results published in the *Census of Manufacturing Industries.*

Organised manufacturing industry has been classified into 63 groups for the purpose of the census following the lines of the United Nations industrial classification, as far as possible. Of these, the first 29 groups cover the major industrial activities in the country and have been included in the annual censuses since 1946. These 29 groups and the number of factories registered under each are as under:

Classification of industry.

S. No.	Industry.	No. of registered factories in existence in 1946.
1.	Wheat flour	66
2.	Rice milling	1,554
3.	Biscuit making	33
4.	Fruit and Vegetable processing ..	19
5.	Sugar	166
6.	Distilleries and breweries ..	46
7.	Starch	5
8.	Vegetable Oils	569

9.	Paints and Varnishes ..	28
10.	Soap	37
11.	Tanning	65
12.	Cement	11
13.	Glass and glassware	126
14.	Ceramics	44
15.	Ply-wood and tea-chests ..	36
16.	Paper and paper board ..	36
17.	Matches	31
18.	Cotton textiles—spinning and weaving ..	483
19.	Woollen textiles	43
20.	Jute textiles	95
21.	Chemicals	174
22.	Aluminium, copper and brass ..	133
23.	Iron and steel	107
24.	Bicycles	5
25.	Sewing Machines	3
26.	Producer gas plants	5
27.	Electric lamps	6
28.	Electric fans	34
29.	General Engineering and Electrical Engineering ..	1,053
		5,013

The prescribed forms have been modelled mostly on the forms used in the United Kingdom and U.S.A., and are of the same pattern for all industries. Every form consists of the following six parts; the first four parts are identical for all industries and the last two parts, which deal with materials used in industry and products made, naturally differ from industry to industry:

1. Part (A)—Name and address of factory, occupier, etc.
2. Part (B)—Capital structure—Paid up and productive capital.
3. Part (C)—Persons employed, man-hours worked and salaries and wages paid.

4. Part (D)—Quantity and value of fuel, electricity, coal gas, lubricating materials and water purchased and consumed.
5. Part (E)—Quantity and value of other materials purchased and consumed.
6. Part (F)—Quantity and value of products and by-products.

The State Governments have appointed Statistics Authorities for administering the Act. For any calendar year, these Authorities are supplied with the prescribed forms by the Directorate of Industrial Statistics, not later than November of the preceding year. The Authorities, in their turn, send three copies of the relevant forms, with a legal notice, to the occupier of each factory before the commencement of the census year. The forms for any calendar year, duly filled in, are due back in the office of each Statistics Authority by the end of February of the following year. They are then scrutinized in that office. Incomplete or discrepant returns are referred back to the occupier concerned for correction. Finally, they are forwarded to the Directorate, with a certificate from the Authority that they are complete and accurate. The Directorate scrutinizes them again and compiles consolidated figures for the different States and the different industries, for publication in the *Census of Manufactures*.

For each of the industries covered by the Census, the Census of Manufactures gives, on State-wise basis, detailed statistical information on the following groups of items:

1. Number of factories.
2. Fixed capital.
3. Working capital.
4. Employment (including number of wage-earners and salary earners, average number of days worked and number of man-hours).
5. Wages and salaries (also, non-cash benefits).
6. Value of materials consumed (including raw materials, fuels, depreciation, etc.).
7. Value of products.
8. Value added by manufacture (item 7 *minus* item 6).

Thus, the Census of Manufactures gives comprehensive information on capital and labour employed, raw materials and fuels used, production, and value added by manufacture. It gives this information for each industry in each State.

Provisional results of the Census of Manufactures covering the year 1950 were published in the July, 1952, issue of the *Monthly Statistics of the Production of Selected Industries of India*. 6,323 factories, engaged in the 28 important industries of India (S. No. 26, Producer Gas Plant in the List of 29 industries on page 663 having been left out), added a total value of Rs. 270.2 crores by way of industrial income. Out of this amount, Rs. 162.3 crores were paid to labour in the form of wages, salaries and other benefits. This amount was distributed among 1.5 million employees. Fixed capital in these industries for the year 1950 was measured to the tune of Rs. 233.4 crores. Rs. 989.7 crores worth of products were produced in these industries. The total material cost was Rs. 719.6 crores.

2,311 factories engaged in the manufacture of durable goods (including iron and steel, machinery, bicycles, non-ferrous metal products, lumber, stone and clay products) employed 315,000 persons, who earned Rs. 39.3 crores during 1950. In these industries the total fixed capital employed was to the tune of Rs. 71.5 crores. The material cost was Rs. 102.9 crores, the value of products Rs. 167.8 crores and the value added by manufacture Rs. 64.8 crores.

4,012 factories engaged in the manufacture of non-durable goods (including manufacturing of food products, alcoholic beverage, textile products, leather, paper, chemicals) employed 1,212,000 persons, who earned Rs. 123 crores during 1950. In these industries the total fixed capital employed was to the tune of Rs. 163.6 crores. The material cost was Rs. 616.6 crores, the value of products Rs. 821.9 crores and the value added by manufacture Rs. 205.3 crores.

For certain individual industries, some of the results of the 1950 census are given below by way of illustration:—

Industry.	No. of factories.	Fixed capital.	No. of employees.	Value added by manufacture.
	Nos.	Lakhs Rs.	Thousands.	Lakhs Rs.
Iron & Steel ..	140	23,14	77	25,10
Bicycles ..	22	61	3	92
Glass & glassware	117	2,95	21	1,78
Biscuit making ..	75	1,26	6	2,08
Woollen Textiles ..	39	1,45	15	4,16

The objectives of a Census of Manufacturing Industries are as under:

1. Estimation of the contribution of manufacturing industries as a whole, and of each unit, to the national income.
2. Systematic study of the structure of industry as a whole, and of each type of industry, and each unit;
3. Analysis of the various factors influencing industry in the country; and
4. Provision of factual and systematic basis for the formulation of Government policy.

The above are the objectives of the Indian Census also. This Census, it should be clearly understood, is only a sub-section of the Census of Production such as that conducted in U.K., U.S.A., Canada, etc. It leaves non-factory industries out, and even among the factory industries it covers only 29 groups, so far. It does not thus cover the medium and small-scale or cottage industries. Steps are, however, said to be well under way to make up this leeway by conducting suitable sample surveys, or otherwise, collecting the requisite data on regular basis.

Monthly Statistics of Industrial Output.

Statistics based on the annual census of manufacturing industries for any year can be available for use, at the earliest,

by the third quarter of the following year. While these detailed statistics are indispensable for long range planning, they become available too late for examining short term behaviour of production and employment in industry. To fill this lacuna, the Directorate of Industrial Statistics are collecting monthly statistics of output on a voluntary basis from the individual undertakings which offer their cooperation, and publishing them, along with estimates, wherever available, of installed capacity of each industry, in the *Monthly Statistics of the Production of Selected Industries of India*, formerly brought out by the Department of Commercial Intelligence and Statistics. Over 90 industries are covered for this purpose, grouped into (A) Mining and Quarrying, (B) Manufacturing and (C) Electric Light and Power. Group (B) covers the largest number of industries and includes, *inter alia*, production of wheat, flour, sugar, salt, *vanaspati*, cigarettes, textiles, footwear, paper, leather, rubber goods, chemicals, glass, cement, stone wares, iron and steel goods, machines, electrical goods, transport equipment. Since the data are collected on voluntary basis, and the cooperating factories do not all furnish them regularly, these statistics of production do not afford a complete picture and have to be used with care.

The table below provides statistics of production of certain industries during the last four years:

Annual Production Statistics.

Industry.	Unit.	1949	1950	1951	1952
Cotton cloth ..	Lakh Yds.	39,048	36,668	40,764	45,984
Sewing machines ..	Nos.	25,032	30,888	44,460	50,040
Electricity generated ..	Lakh kwts.	49,090	50,880	58,517	61,250
Radio Receivers ..	Nos.	16,836	44,340	68,100	71,496
Cement ..	'000 tons.	2,102	2,612	3,196	3,538
Salt ..	'000 mds.	55,620	71,316	74,376	76,860
Vegetable Oil products ..	Tons.	155,544	171,636	172,320	190,812

All the industries listed above recorded progressively rising production during the four-year period. The table gives some idea of the pace of industrialization of the country since 1949.

In this monthly bulletin are also published the Index Numbers of Industrial Production, which will be dealt with in the next chapter.

Other industrial statistics.

The publication entitled *Large Industrial Establishments in India,* issued by the Labour Bureau, Ministry of Labour, contains a complete list of factories in the various States in India employing 20 workers or more. The list is arranged by industry and gives details of location of the factory and size, according to average daily employment. The industries covered include textile engineering, minerals and metals, food, drink and tobacco, paper and printing, lumber, stone and glass manufactures, gins and presses, etc.

The Monthly Statistics of Cotton Spinning and Weaving in Indian Mills issued by the Directorate of Industrial Statistics, gives the relevant data on Cotton Spinning and Weaving in India collected in accordance with the requirements of the Cotton Industry (Statistics) Act of 1926. These data relate to production of different classes of cotton goods, yarn spun, cotton pressed, and consumption of cotton in Indian mills.

The Monthly Coal Bulletin, issued by Chief Inspector of Mines, presents a monthly picture of statistics of the coal industry, such as output of coal, labour employed in collieries, despatches and stocks, etc.

Collection of Statistics Act, 1953.

The Government of India had very limited legal powers to collect statistics for the various subjects allocated to the Centre under the constitution, particularly in the industrial and commercial field. The existing arrangements for obtaining trade statistics through the Customs authorities and statistical data in respect of specific commodities under the relevant Acts, were also not very comprehensive. The Government had, accordingly, to depend on the voluntary cooperation of the interests

concerned in regard to the collection of statistics from industrial concerns and business houses. This cooperation did not always come readily; the lack of it became most marked when, in order to assess the extent to which foreign owned and controlled firms engaged Indians and foreign nationals, they issued a notification in 1952 calling upon undertakings to furnish the relevant statistics, and found that the response to the notification was not very satisfactory. The Government felt it expedient to arm themselves with powers to compel all concerns engaged in (*a*) trade and commerce with foreign countries, (*b*) inter-State trade and commerce, (*c*) trading firms or corporations, incorporated, registered or otherwise permitted in India, including banking, insurance and other financial corporations, and (*d*) stock exchanges, to furnish such information. Accordingly, they moved a Collection of Statistics Bill which was passed by both the Houses of Parliament in August, 1953. The Act, while replacing the Industrial Statistics Act, 1942, incorporates all its existing provisions and provides, in addition, for the collection of information and statistics in respect of matters relating to (*a*), (*b*), (*c*) and (*d*) mentioned above. It gives very wide powers to the Government of India to collect statistics; Section 3 in it empowers them to call for data relating to any of the following matters :

(a) any matter relating to any industry or class of industries;

(*b*) any matter relating to any commercial or industrial concern or class of commercial or industrial concerns, and in particular, any matter relating to factories;

(*c*) prices of commodities, attendance, wages, hours of work, employment, un-employment, disputes, etc., so far as they relate to welfare of labour and conditions of labour.

This is the first comprehensive legislation in India to facilitate the collection of statistics relating to industries, trade and commerce and is, as such, very welcome. It opens a new chapter in regard to the collection of industrial, commercial and labour statistics in India.

WAGES

Wages statistics can be studied under two broad heads: agricultural and industrial.

AGRICULTURAL WAGES

Until recently, the data available on the duration of agricultural employment, wage rates and earnings were meagre. The available material comprised either (*a*) periodical returns of 'average wages per month' of skilled and unskilled labour which were published twice a year, for each district, in the State Gazettes and annually, for the selected districts, in the *Prices and Wages*, or (*b*) data collected through socio-economic rural enquiries. In 1905, the Government of India decided to discontinue the half-yearly wage returns and asked the State Governments to conduct a five-yearly wage census. Madras, U.P. and Punjab conducted quinquennial censuses of rural and urban wages and Bombay and Madhya Pradesh annual censuses. The data collected were not very reliable; the collecting agency, in addition to its onerous duties, could pay only perfunctory attention to the collection of wages data. The proforma was revised and made more comprehensive in 1919, whereafter the quality and content of the statistics considerably improved. Some defects, however, still remained: the data were not collected in all the States; and, wherever they were collected, the collection was not done simultaneously or on a uniform basis.

Data published by Directorate of Economics & Statistics.

With a view to ensuring the collection of statistics of wages of agricultural labour on a uniform and regular basis, this Directorate under the Ministry of Food and Agriculture finalized a scheme and the necessary proforma in consultation with the State Governments in 1950. The salient features of the scheme are:—

(1) Agricultural labourers are divided into the following heads:

I. Skilled labour,

(*a*) Carpenters,

(*b*) Black-smiths,

(*c*) Cobblers;

II. Field Labour;
III. Other Agricultural Labour;
IV. Herdsmen.

(2) For groups II, III and IV above, which constitute unskilled labour, wages are given separately for man, woman and child.

(3) Wages relate to labour employed on a daily wage rate and paid either in cash or in kind, the latter being converted into cash equivalent.

(4) The data are collected on monthly basis and the wages are those most commonly current during the month.

(5) The data are collected from each district; at any rate, from one in every five. In each district, one village is selected so as to represent the wages and general agricultural conditions of the district.

(6) The district returns are consolidated at the State headquarters, and a consolidated return furnished to the Directorate.

The first issue of the *Indian Agricultural Wages Statistics*, 1950 gave the data collected according to the above-mentioned scheme from 14 States. The number of reporting States increased in the subsequent years. The tables in this publication give the State-wise maximum and minimum wages, together with the months of the year when they were prevailing in the case of each type of labour. They thus give a fair idea of the busy and slack seasons in respect of each category of labour.

Agricultural Labour Enquiry (ALE).

In pursuance of the recommendation of the Tripartite Labour Conference held in 1943, the Government of India decided to conduct a sample enquiry into the conditions of agricultural workers to determine the proportion of hired labour employed in agriculture, the wages paid, concessions allowed, etc. Before, however, the enquiry could be conducted, the Minimum Wages Act, 1948, was passed, under which minimum rates of wages for employees in a number of scheduled employments, both agricultural and non-agricultural,

were to be fixed within a specified period. Before such wages could be determined, an all-India enquiry into agricultural workers' living conditions was considered necessary, and this was started, in close collaboration with the State Governments, in the middle of 1949.

The object of ALE was to collect data on employment, earnings, cost and standard of living and indebtedness of agricultural workers with a view to considering what protective and ameliorative measures, including the fixation of minimum wages under the Act of 1948, should be undertaken. It was conducted in 813 villages in the country selected on the principle of stratified random sampling. For the purpose of the enquiry, the country was divided into 23 units, each under the charge of a Superviser. The primary data were collected with the help of 52 Investigators and 212 Deputy Investigators. The period of the enquiry was a year and the data collected also referred to the same period.

The survey was conducted by the Central Ministry of Labour, in three different stages through three different schedules: 1. General Village Schedule, 2. General Family Schedule and 3. Intensive Family Schedule. The main object of the first two stages was to get an idea of the general economic conditions and the employment structure in the village and to delimit the agricultural labour families for the purpose of the Intensive Family Survey, which formed the third stage. An agricultural labour family was regarded as one in which either the head of the family or 50% or more of the earners reported agricultural labour as their main occupation. For the first two stages, the sampling unit was a village and for the third stage, it was the agricultural labour family.

The items on which data were collected through the General Village Schedule included land tenure, population, number of agricultural and non-agricultural families, area under crops, wage rates, prices, in the 813 selected villages. In the General Family Schedule information was collected for all the families in each selected village, regarding the size and earning strength of each family, occupational structure, holdings, implements, livestock and housing. Out of the list of agricultural labour families emerging from the General Family Schedules, a sample of agricultural labour families was selected

and a monthly Intensive Survey thereof was conducted to collect data on employment, unemployment, wage rates, earnings, indebtedness, cost and standard of living.

ALE reports are being brought out gradually. Already, separate reports in respect of a village each in Madras, Bihar, West Bengal, Madhya Pradesh, Orissa, Assam, U.P. and Mysore have been published. A publication entitled *Agricultural Wages in India, Volume I,* was issued in 1952. It throws light, *inter alia,* on the agricultural conditions in different States, quantitative analysis of agricultural labour families, wages and conditions of employment, prices, *begar,* (involuntary labour), etc. The following table, extracted from the *Agricultural Wages in India, Volume I,* gives the average daily wage rates for men, obtaining in the sample villages in the Part A States, taking into consideration all agricultural operations and the different modes of wage payment (excluding piece rates):

Average Daily Wage.

STATE	Rs.	As.	P.
Assam	2	0	8
Punjab	2	0	1
West Bengal	1	12	0
U.P.	1	3	8
Bihar	1	2	9
Bombay	1	1	6
Madras	0	15	7
Madhya Pradesh	0	12	1
Orissa	0	12	0

The above table has been arranged in descending order and shows that the agricultural wages paid in Orissa are the lowest among the Part 'A' States.

INDUSTRIAL WAGES

Wages relating to mining are regularly published in the *Annual Reports* of the Chief Inspector of Mines and those relating to tea plantation labour in the *Indian Tea Statistics,* the latter being brought out by the Directorate of Economics and Statistics of the Ministry of Food and Agriculture. The *Annual Reports* on the working of the Indian Factories Act in

the States contain some data regarding the wages of skilled and unskilled workers in factories. Also the *Annual Reports* on the working of the Indian Trade Unions Act, 1926 and the Workmen's Compensation Act, 1923, throw some light on the subject, besides the *Labour Gazettes* of the States. *Ad hoc* enquiries into family budgets and cost and standard of living of plantation workers conducted by the Labour Bureau, as also the reports of the various Labour Enquiry Committees give some figures. A mass of uncoordinated information is thus available, which is unsuitable for time study.

Recently, however, the Labour Bureau has started compiling an Index Number of Earnings of Factory Workers in India* on the basis of the statistics of earnings of factory workers collected under the provisions of the Payment of Wages Act, 1936. This index fills up a long felt want. For some time, the *Indian Labour Gazette*, brought out by the Labour Bureau, has also been publishing data on the following, among other items:

1. Employment in factories, Central Government, Cotton mill industry and coal mines;
2. Employment exchange statistics;
3. Wages and earnings
   (*a*) Total and average earnings of factory workers drawing less than Rs. 200 per month, excluding railway employees;
   (*b*) Minimum wages in cotton mill industry;
   (*c*) Average weekly earnings of coal mining labour.

The table below gives the minimum wages in the cotton mill industry in certain industrial centres:

	Bombay			Nagpur			Madras			Kanpur		
	Rs.	A.	P.	Rs.	A.	P.	Rs.	A.	P.	Rs.	A.	P.
Minimum Basic Wages ..	30	0	0	26	0	0	26	0	0	30	0	0
Dearness Allowance												
—January, 1952 ..	58	2	0	43	9	8	44	7	0	55	0	0
—January, 1953 ..	62	0	0	41	9	2†	42	9	0	56	14	0

† For November, 1952.

* For its construction, see next Chapter.

PRICES

Statistics of Prices of Commodities in India can be found in the following official publications, besides various non-official bulletins, journals, market reports, etc.—

1. *Prices and Wages;*
2. *Wholesale Prices of Certain Selected Articles of Trade at Selected Stations in India;*
3. *Indian Trade Journal;*
4. *Monthly Survey of Business Conditions in India;*
5. *Index Number of Wholesale Prices in India;*
6. *The Journal of Industry and Trade;*
7. *Agricultural Situation in India;*
8. *Bulletin of Agricultural Prices;*
9. *Indian Agricultural Prices Statistics;*
10. *Indian Agricultural Statistics;*
11. *Season and Crop Reports;*
12. *Indian Food Statistics;*
13. *Bulletin on Food Statistics;*
14. *Statistical Bulletins issued by the States;*
15. *Reserve Bank of India Bulletin;*
16. *Weekly Bulletin of Statistics;*
17. *Monthly Bulletin of Statistics.*

Prices and Wages, State *Gazette* and the *Government of India Gazette* for long constituted the only source for prices statistics. They related generally to raw materials. These data were, however, not quite trustworthy since the revenue agency, which compiled them, did not devote sufficient time and care in collecting them, did not wholesale prices were often confused with retail ones and old quotations were repeated although the market prices had changed. The *Indian Trade Journal* and *Monthly Statement of Wholesale Prices* (No. 2 in the above list), both publications of the Department of Commercial Intelligence and Statistics, started giving publicity to prices of agricultural and manufactured goods collected from a variety of sources, including non-official agencies. The *Monthly Survey of Business Conditions in India* issued by the office of the Economic Adviser, Ministry of Commerce and Industry, published, *inter alia,* a review of prices of cotton and jute manufactures, iron and steel, sugar, coal, foodgrains, oil-

seeds, raw cotton and jute and tea. This publication was merged, in the middle of 1951, with *The Journal of Industry and Trade*, a monthly, issued by the Ministry of Commerce and Industry, which reproduces the prices of individual commodities—both agricultural and non-agricultural—forming the basis of Economic Adviser's index of wholesale prices. These prices are also issued by the latter in a weekly bulletin entitled *Index Number of Wholesale Prices in India* (*vide* next chapter for the commodities included in this index). For commodities under Government control, controlled prices are generally used in this index, and, therefore, the two publications just cited (Nos. 5 and 6 in the above list) give open market prices for uncontrolled commodities and controlled prices for those under Government control.

The *Agricultural Situation in India*, a monthly journal issued by the Directorate of Economics and Statistics under the Ministry of Food and Agriculture, publishes prices of sugarcane (*a*) delivered at factory gates and (*b*) actually received by growers, the data being collected by the Indian Institute of Sugar Technology, Kanpur. It gives the pool auction prices of coffee, obtained from the Indian Coffee Board. Besides, it gives (*i*) wholesale prices of certain important agricultural commodities and animal husbandry products at selected centres in India, (*ii*) wholesale ration rates of foodgrains, (*iii*) prices of certain important agricultural commodities in foreign markets, (*iv*) retail prices of fruits and vegetables, (*v*) retail prices of livestock and livestock products, (*vi*) retail prices of fish, eggs and poultry, besides index numbers of wholesale and retail prices. The *Bulletin of Agricultural Prices*, issued by the same Directorate, gives prices data on some of the above mentioned items; these data become available to the readers weekly. The *Indian Agricultural Prices Statistics*, issued annually by the same Directorate, gives, *inter alia*, (*i*) farm (harvest) prices, (*ii*) procurement prices for foodgrains, (*iii*) maximum wholesale selling prices of foodgrains, (*iv*) wholesale market prices, (*v*) retail ration rates of foodgrains, (*vi*) retail market prices, besides Economic Adviser's index numbers. The data are given on State-wise basis for important assembling and consuming towns and for a very large number of agricultural commodities, raw as well as processed.

Harvest Prices.

Mention has been made above of the farm (harvest) prices of agricultural commodities. Farm price, in strict theory is the wholesale price received by the farmer for his produce at the farm at the harvest time. The farm (harvest) prices, available in India, do not, however, satisfy this definition. Their definition and method of collection have all these years been governed by local custom and usage and, therefore, differ from State to State. For instance, for Assam, they are the wholesale prices in four markets during the harvest time, for Bombay retail prices, for Orissa, West Bengal and Madras they are retail prices in 2 or 3 principal markets (for oilseeds and rice in Madras they are wholesale prices), only in Punjab they are the prices received by farmers in selected villages of each Kanungo circle. Such harvest prices have been collected in what are now known as Part A States since the last quarter of the 19th Century as a part of the Returns of Agricultural Statistics by the *Patwaris*. Upto 1946-47, they were published in the *Indian Agricultural Statistics* brought out by the Directorate of Economics and Statistics; subsequently, with the integration of this Directorate's publications, all prices data were transferred to a new publication, *Indian Agricultural Prices Statistics*, referred to above, and this publication now contains them. These data, as said above, lack uniformity and are not harvest prices in the strict sense of the term. To remove these defects, the Directorate finalised a scheme in consultation with the State Governments in 1950. This scheme defines farm (harvest) price as the average wholesale price at which the commodity is disposed of by the producer to the trader at the village site during the specified harvest period. Under the scheme, the data are collected from representative villages in each district, on every Friday, for the most common variety; simple arithmetic average of the prices of villages in a district constitutes the district average; these district averages, weighted by the proportion of quantity produced, give the State average of farm (harvest) prices. Some of the States have already started collecting data on the lines of this scheme from 1950-51; in course of time, therefore, farm (harvest) prices in India will become available on not only a uniform basis but also in terms of a scientific concept. They

will be available not only for the Part A States but also for others. The *Season and Crop Reports* of the States will also contain them.

There is another series of harvest prices too. The Department of Commercial Intellligence and Statistics used to collect from the branches of the Imperial Bank of India, the prices of agricultural commodities for a period of about eight weeks after the arrival of harvest in the market. These prices were published in the name of harvest prices in the *Indian Trade Journal* by that Department; since the transference of all work relating to agricultural statistics from that Department to the Directorate in 1948, they have been published by the latter in the *Agricultural Situation in India*, and, in order to depict the exact nature of these prices more precisely, they have been renamed harvest-season prices.

Controlled Prices.

Since the inception of food controls in the country in 1943-44, prices of foodgrains have been controlled at various stages. There are the procurement prices at which Government obtain foodgrains from the producers or traders, wholesale issue prices at which Government supply foodgrains to ration-depots, retail issue prices at which ration-depots sell to consumers, and maximum wholesale selling prices or ceiling prices beyond which the market prices are not expected to go in unrationed areas. These data were first published in the *Manual of Food Statistics*, 1946, by the then Ministry of Food and later in publications No. 12 and 13 mentioned in the list on page 675, brought out by the Directorate of Economics and Statistics.

Controlled prices for sugar, cotton, fertilizers etc. are published in the *Agricultural Situation in India* and those of textiles, cement, iron and steel, coal, petrol etc. in the weekly bulletin relating to E.A's wholesale prices index.

State Bulletins.

The *Statistical Bulletins* brought out by most of the State Governments also contain detailed prices statistics, in many cases on district-wise basis.

Bullion and Securities.

The *Reserve Bank of India Bulletin* publishes prices of gold and silver, and of Government of India securities and industrial securities. The *Weekly Bulletin of Statistics* and the *Monthly Abstract of Statistics* both brought out by the Central Statistical Organisation also contain prices of gold and silver.

Improvements Effected.

From the above account, it will be seen that the Office of the Economic Adviser and the Directorate of Economics and Statistics are the two chief offices under the Government of India for the collection and publication of prices statistics. These offices obtain their data largely from the State Governments or other Government sources like the Collectors of Customs, though in some cases they get them from non-official bodies too, such as the Imperial Bank of India, Chambers of Commerce, trading firms, etc. In the States, considerable attention has, in recent years, been paid to improving the quality and content of these statistics, thanks to the need for reliable data for administering various controls. In a number of them, the revenue agency, which was once the only price-reporting agency, has now been replaced, in so far as price-reporting is concerned, by Economic Intelligence Department, State Marketing Department, Statistical Bureau, Procurement Department, Civil Supplies Department, etc.* Officials of these departments are posted in the important marketing centres in almost every district; it is their duty to keep in touch with the market conditions and to keep the State headquarters, as well as the specified offices of the Central Government, posted with market quotations periodically. In order that these quotations may be obtained by the Government of India on a uniform and comparable basis from different parts of the country, they have laid down uniform instructions for the States. For example, the connotation of the term price so far as the Directorate of Economics and Statistics is concerned is the wholesale, closing market quotation, for sale ex-godown, for the naked grain (that

---

* The Government of India have since appointed an Agricultural Prices Enquiry Committee, in November 1953, to examine the existing machinery for the reporting of prices of agricultural commodities and to suggest improvements in the existing machinery.

is, exclusive of cost of bag and incidental charges), for coarse variety of the commodity and for Saturdays. Thus, the quality of prices statistics has considerably improved now.

Much progress has also been made in the matter of dissemination of market intelligence. The Agricultural Marketing Adviser to the Government of India initiated, during World War II, a market intelligence service which included broadcast of a weekly market report by the all-India radio and of daily market quotations for Hapur. This service still continues; it has only been transferred to the Directorate of Economics and Statistics. The weekly bulletins of the Directorate and the Office of the Economic Adviser make the data available to the public with a very little time lag. The State Governments' bulletins and radio broadcasts do likewise.

Suggestions for Further Improvement.

There is still some leeway to be made up. Firstly, in the absence of standardization of qualities of commodities, it is at present not quite possible to ensure strict comparability of quotations for all parts of the country. Secondly, information on price-spreads is not available on a regular basis; *Reports on Marketing Surveys* issued by the Agricultural Marketing Adviser to the Ministry of Food and Agriculture no doubt contain some information on margins of profits of middlemen between the producer and ultimate consumer, but these reports are now pretty old while market practices have undergone considerable changes in many parts of the country. According to *Wheat Marketing Survey,* for instance, in some cases the cultivator got about Re. 0-9-3 out of each rupee paid by the consumer; but this information relates to the pre-War period when there were no price controls or movement restrictions, or when regulated markets had not been established. There is need for collecting and keeping up-to-date information on price-spreads. Thirdly, the huge mass of data being collected in each State and also being compiled by the Central Government need to be processed so as to enable study of price trends. They may, for instance, be utilised for constructing indices which would show the variation over a period of time. An account of what is being done, and what more needs to be done in this direction will be found in the next chapter. Lastly,

there is need to so reform the collection of data as to enable the construction of index numbers of prices paid by farmers for the purchases of their requirements and prices received by them for their produce. This study is made in several foreign countries including U.S.A.; it gives a measure of agricultural prosperity and serves as a basis for the country's price policy.

TRADE STATISTICS

The Department of Commercial Intelligence and Statistics is, and has been, the sole organisation under the Government of India for the collection, compilation and publication of India's foreign and inland trade statistics.

Monthly Foreign Trade Statistics.

The basic statistics of foreign trade of India were formerly published in the *Accounts relating to the Foreign Sea and Air-borne Trade and Navigation of India* and *Accounts relating to the Trade of India by Land with Foreign Countries.* It was recently considered desirable to publish the entire statistics of foreign trade by sea, air and land in a single publication. From April, 1952, this proposal was given effect to, partially: the statistics of India's overland trade with Pakistan, Afghanistan, Iran and Burma were included in the *Foreign Sea and Air-borne Trade Accounts* and the publication designated as *Accounts relating to the Foreign (Sea, Air and Land) Trade and Navigation of India.* This inclusion does not amount to complete amalgamation of overland trade figures with sea and air-borne trade figures, as the commodity classifications for the two sets of figures are not yet uniform. Foreign trade by land, therefore, constitutes only a section in the publication, separate from the section on foreign trade by sea and air. It has also not been possible to include in the new publication the statistics of India's trade with Nepal, Tibet, Sikkim and Bhutan which relate to the traffic in selected articles, in quantity only, registered at certain selected railway stations adjacent to India's land frontier routes to those countries. These statistics are being published, on a monthly basis, in the *Indian Trade Journal.*

The *Accounts relating to the Foreign (Sea, Air and Land) Trade and Navigation of India* is a monthly publication and gives data on the following, among other items, for the month

concerned of the current year and previous two years, as also progressive totals from April for three years:—

Part A—Foreign Trade by Sea and Air—

1. Imports,
2. Re-exports,
3. Exports,
4. Trade in Treasure.

Part B—Shipping-vessels in the Foreign Trade.

Part C—Foreign Trade by Land—

1. Imports,
2. Re-exports,
3. Exports.

For each Part, the publication gives summary tables and detailed tables.

Foreign Trade by Sea and Air.

The trade dealt with in the *Accounts* is the trade by sea and air registered at Indian ports (sea-ports and/or air-ports open for foreign traffic) between those ports and foreign countries. Import figures show generally the trade as declared by importers on the bills of entry passed through the Customs Houses upto the last working day of the month. According to the revised procedure for the registration of export trade statistics adopted in the Customs Houses since 1951-52, these statistics are now being compiled on the basis of the trade as declared in the shipping billls passed upto the last working day of the month, with necessary adjustments on account of short shipments.

Exports are divided into (*a*) exports of Indian merchandise and (*b*) re-exports of foreign merchandise, the distinction being made in the Shipping Bills from which trade statistics are compiled.

The articles in the detailed tables are arranged in alphabetical order. The following are the heads of classification:—

I. Food, drink and tobacco,

II. Raw materials and produce and articles mainly unmanufactured,

III. Articles wholly or mainly manufactured,

IV. Living Animals,

V. Postal articles.

The figures of quantities and values are those declared by importers and exporters in Bills of Entry and Shipping Bills, as subsequently checked by Customs officials. The weight recorded is the net weight, that is, it is exclusive of the weight of the covers or packing. With effect from 1951-52, the basis of export valuation conforms strictly to f.o.b. basis inclusive of export duty and for cess when leviable.

The *Accounts* do not include complete figures of all imports on Government account owing to a special system of clearance of Government Stores. The figures are, however, adjusted later when the wanting particulars are collected by Customs Houses. In so far as foodgrains are concerned, approximate total value of such residuary imports on Government account as are not included in the main tables, are given separately with effect from April 1952. The *Accounts* do not also include transhipment trade. As for transit trade, trade of foreign countries situated on the border of India passing through Indian ports in transit is included in the import and re-export tables; it is also being recorded separately since April, 1949. The transit trade of Pakistan with overseas countries passing through Indian ports has been excluded from the import and re-export tables with effect from April, 1952.

Shipping.

The tables relating to the movements of shipping included in the *Accounts* relate only to entrances and clearances "with cargo", that is, vessels which arrive at a port or ports in India for discharging cargo from abroad or which depart from a port or ports in India after loading their cargo for discharge at a destination abroad. The following classes of vessels are distinguished in the shipping tables:—

1. Indian,
2. British,
3. Country Craft,
4. Foreign.

The tonnage shown in the tables represents the net registered tonnage of vessels.

Arrivals and departures of ships "in ballast," that is, without cargo, are recorded separately and published in Volume II

of the *Annual Statement of the Foreign Sea-Borne Trade of India.*

The Government of India appointed a Port and Shipping Statistics Committee in 1953, with Shri M. A. Master as the Chairman, to examine the present system of collection, maintenance and publication of port and shipping statistics in India, and, having regard to the new needs of Government agencies, to suggest what further statistics should be collected, by what agency they should be collected and how they should be published. In August 1953, the Committee sent out questionnaires to the various ports and customs administrations in India. The report* of the Committee is awaited.

Foreign Trade by Land.

As already mentioned above, these statistics are now available in the *Accounts* for Pakistan, Afghanistan, Iran and Burma, while for other countries they are available in the *Indian Trade Journal.*

Balance of Trade.

The table below illustrates how the 'balance of trade in merchandise' and the 'visible balance of trade' are struck:—

Overall Balance of Trade—April, 1953.

(Lakhs of Rs.)

	By Sea and air	By land	Total
1. *Trade in Merchandise*—			
(*i*) Exports ..	37,93	37	38,30
(*ii*) Re-exports ..	51	..	51
Less transit trade ..	2	..	2
Net Re-exports ..	49	..	49

TOTAL EXPORTS ..	+38,42	+37	+38,79
(*iii*) Imports ..	56,77	71	57,48
Less transit trade ..	2	..	2
NET IMPORTS ..	−56,75	−71	−57,46
Balance of Trade in Merchandise ..	−18,33	−34	−18,67
2. *Trade in Gold & Silver—*			
(*i*) Exports and Re-exports ..	..	..	..
(*ii*) Imports ..	1	..	..
Balance of Trade in Gold and Silver ..	−1	..	−1
3. *Total overall Visible Balance of Trade* ..	−18,34	−34	−18,68

Annual Foreign Trade Statistics.

Full particulars of the import and export trade in each article and with each foreign country are published in two volumes on the *Annual Statement of the Foreign Sea-borne Trade of India.* Volume I gives abstract and detailed tables of imports and exports, while Volume II those of trade and shipping with each country and at each port. This publication thus affords very comprehensive information on the volume, value and direction of India's foreign trade in each article.

Other Foreign Trade and allied Publications.

Monthly statistics of India's foreign trade by sea, air and land with certain principal countries, and also classified according to certain groups of articles by different currency (hard currency, medium currency and soft currency) areas are pub-

lished in the monthly *Statistics of Foreign Sea-borne Trade of India by Countries and Currency Areas.*

Monthly statistics of gross collection of customs duties (import and export), which were formerly published in the *Accounts relating to the Foreign, Sea and Air-borne Trade and Navigation of India* are now published in the monthly *Customs and Excise Revenue Statement of the Indian Union.*

Other publications on certain aspects of foreign trade include:

1. *Indo-Pakistan Trade Statistics* (monthly),
2. *Exports of Indian Artware and Sports Goods* (monthly),
3. *Accounts relating to the Coasting Trade and Navigation of India* (monthly),
4. *Indian Customs Tariff* (annual).

Inland Trade Statistics.

Statistics of movement of goods from one State to another were being published by the then Department of Statistics (DCIS) until 1922 when they were discont nued on grounds of economy. Their publication has since been revived in the monthly *Accounts relating to the Inland (Rail and River-borne) Trade of India.* This publication gives a summary view of the total quantity of merchandise exported from and imported into each block, as also the quantities of each commodity so moved. It presents the data in a two-way table, with the names of the same blocks arranged horizontally as well as vertically, so that exports from one block on the vertical side to each of the blocks on the horizontal side can be given. Generally each State forms a block; chief port or ports constitute a separate block from the State in which they are situated. From this publication, therefore, movement of merchandise can be studied on the following lines:—

(1) Trade of a State with other States
(2) Trade of a port with other ports
(3) Trade of a State with a port.

Trade within the same block is excluded. The commodities included in the publication have been classified into over 30 groups, which include animals and livestock, cement, cotton twist and yarn, cotton piecegoods, fruits, grains, pulses and

flour, hides and skins, raw jute, gunny bags and cloth, iron and steel goods, lac and shellac, oil cakes, oils, oilseeds, ghee, salt, sugar, wood and timber, raw wool.

The quantities in this publication refer to those carried by railways and steamer services as recorded in invoices, and are exclusive of cover or packing material. Values are not recorded by these services and are, therefore, not given in this publication.

Raw Cotton Trade Statistics.

Statistics of movement of raw cotton from one State to another are given in the monthly *Raw Cotton Trade Statistics (Rail and River-borne)*. The data relate to all qualities and varieties of cotton.

Retail Trade.

While fairly detailed and reliable statistics are available in respect of India's foreign and inland trade, no data are available regarding the volume of retail trade. This is a long-standing gap in our trade statistics. Now that sales tax has been introduced in a number of States in the country, the machinery which administers this tax can be used for collecting statistics of retail trading. Such statistics are published for U.K. by the Board of Trade.

Review of Trade.

The Office of the Economic Adviser, under the Ministry of Commerce and Industry, publishes the annual *Review of the Trade of India*. This publicat on was formerly issued by the Department of Commercial Intelligence and Statistics. Besides reviewing the general economic trends, it gives an analysis, together with statistical tables, of the foreign trade, direction of overseas trade, coastal trade, balance of trade, etc.

FINANCIAL STATISTICS

Financial statistics can be broadly divided into two heads: (1) Public Finance and (2) Banking, Currency and Exchange.

PUBLIC FINANCE

Statistics relating to public finance may be classified as those relating to (*a*) public revenue and (*b*) public expenditure. They are available separately for the States and the Central Government, and in respect of the latter, separately for the Railway Budget and the General Budget. These statistics are by far the most comprehensive and up-to-date, and, as the principles underlying their compilation and presentation are almost similar, we will take up only the General Budget of the Government of India for describing these principles.

Central Budget.

The *Budget of the Central Government* for any financial year (April to March) is laid before the Parliament towards the end of February of the preceding financial year. It contains the following statements:—

I. General Statement of the Revenue and Expenditure met from Revenue of the Central Government;
II. General Statement of the Receipts and Disbursements of the Central Government;
III. Detailed Statement on Revenue;
IV. Detailed Statement on Expenditure;
V. Detailed Statement on Receipts and Disbursements.

According to Article 266 of the Constitution of India, revenues received by the Government of India, loans raised by her by the issue of treasury bills, loans or ways and means advances and moneys received in repayment of loans are to form one consolidated fund called the 'Consolidated Fund of India'. All other public moneys received by or on behalf of the Government of India are credited to the 'Public Account of India'. Parliament may also establish a contingency fund in the nature of an imprest called 'Contingency Fund of India' out of which the President may make advances for the purposes of meeting unforeseen expenditure, pending authorisation of such expenditure by Parliament.

As its very name implies, General Statement I of the *Budget of the Central Government* gives statistics of revenue going into the 'Consolidated Fund of India', and of expenditure incurred out of that Fund. General Statement II gives statis-

tics of receipts into the Consolidated Fund, Contingency Fund and Public Account, and of expenditure out of each and shows the opening and closing cash balances of the Government. Details follow in the subsequent three statements. Budget estimates are given for the ensuring year, both the Budget and Revised estimates for the current year, and Accounts for the immediately preceding year.

Revenue and Expenditure.

The table below summarises General Statement I of the *Budget of the Central Government for 1953-54:*

(Crores of Rs.)

	Accounts 1951-52	Budget Estimate 1952-53	Revised Estimate 1952-53	Budget Estimate 1953-54
Total Revenue ..	515.4	405.0	418.6	439.3
Total Expenditure met from Revenue ..	387.3	401.3	422.4	438.8
Surplus (+) Deficit (—) ..	(+)128.1	(+) 3.7	(—) 3.8	(+) 0.5

The revised estimates for 1952-53 showed a deficit of Rs. 3.8 crores as against a surplus of Rs. 3.7 crores in the budget estimates; the revised estimates of revenue showed an improvement of over Rs. 13.5 crores as compared with the budget estimates due to better receipts under customs and income tax; the revised estimates of expenditure met from revenue showed an increase of over Rs. 21 crores over the budgetted figure, owing mainly to additional payments required to be made to the States under the Finance Commission's recommendations.

Of the estimated expenditure of Rs. 438.8 crores for 1953-54, Rs. 199.8 crores was to be incurred on Defence Services and the balance of Rs. 239 crores on the Civil side. For the latter, the expenditure in 1953-54 was estimated to increase by Rs. 34.5 crores over last year, due mainly to larger provision

for several development items, like basic and social education, development of handloom industry, industrial housing, backward classes, National Extension Organisation, community development projects, etc.

Out of the various heads of Central Revenue and Expenditure, summary figures for which are given in the above table, the principal ones are given below:

Revenue

1. Customs.
2. Union Excise Duties.
3. Corporation Tax.
4. Income Tax.
5. Opium.
6. Contribution from Railways
7. Posts and Telegraph.
8. Currency and Mint.
9. Civil Works.

Expenditure

1. Direct Demands on the Revenue.
2. Irrigation.
3. Debt Services.
4. Civil Administration.
5. Civil Works.
6. Defence Services.
7. Contributions and Miscellaneous Adjustments between Union and State Governments.

Receipts and Disbursements.

The table below summarises General Statement II of the *Budget of the Central Government for 1953-54:*

(Crores of Rs.)

	Accounts 1951-52	Budget Estimate 1952-53	Revised Estimate 1952-53	Budget Estimate 1953-54
Opening Balance ..	161.78	158.68	162.69	79.84
Total Receipts ..	225.38	94.39	106.58	174.45
TOTAL ..	387.16	253.07	269.27	254.29
Total Disbursements ..	224.47	169.99	189.43	202.60
Closing Balance ..	162.69	83.08	79.84	51.69

It will be seen that the Budget for 1952-53 provided for an overall deficit of Rs. 76 crores to be met from the opening cash

balance of Rs. 159 crores, leaving, at the end of the year, a closing balance of Rs. 83 crores. The revised estimates showed an increase in the overall deficit from Rs. 76 crores to Rs. 83 crores, leaving a closing balance of Rs. 80 crores.

General Statement II opens with the Revenue Deficit or Surplus arrived at in General Statement I, and then gives receipts which are not of revenue character and disbursements not met from revenue. The heads for which receipts and disbursements are given in this Statement are almost similar and are as under:

I. CONSOLIDATED FUND OF INDIA.

1. Revenue Surplus/Deficit as per General Statement I.
2. Capital Account outside the Revenue Account.
3. Permanent Debt.
4. Floating Debt.
5. Loans and Advances by the Central Government.

II. CONTINGENCY FUND OF INDIA.

III. PUBLIC ACCOUNT.

1. Unfunded Debt.
2. Depreciation and other Reserve Funds.
3. Appropriation for Reduction or Avoidance of Debt.
4. Other Deposits and Advances.
5. Remittances.
6. Transfer of Cash between England and India.

IV. OPENING AND CLOSING BALANCES.

Apart from the *Budget* of the Central and State Governments, statistics on different aspects of public finance are also published in the *Combined Finance and Revenue Accounts* issued by the Ministry of Finance, Government of India. They are also reproduced in the *Statistical Abstract,* publications of the Reserve Bank and elsewhere. They arise from the working of Government administration. As such, there is a regular flow of these statistics from the different departments of the Government to the Finance Department, who consolidate them for presentation to the Legislative bodies in the case of States and Parliament in the case of the Centre.

The *Report on Currency and Finance,* issued annually by the Reserve Bank of India, devotes a chapter to Public Finance, wherein is given an analytical account, supported by detailed statistics, of the finances of the Central and State Governments, Railway Finance, and Public Debt. The *Reserve Bank of India Bulletin* gives monthly statistics of revenue, expenditure and public debt of the Central Government, Government of India Treasury Bills offered, tendered, sold and outstanding, and deposits on account of small savings.

BANKING, CURRENCY AND EXCHANGE

The chief sources of these statistics are as under:

(1) *Reserve Bank of India—Statement of Affairs* (Weekly);

(2) *Scheduled Banks—Statement of Affairs* (Weekly);

(3) *Reserve Bank of India Bulletin* (Monthly);

(4) *Report on the Trend and Progress of Banking in India* (Annual);

(5) *Statistical Tables relating to Banks in India* (Annual);

(6) *Statistical statements relating to the Cooperative Movement in India* (Annual);

(7) *Report on Currency and Finance* (Annual).

All these publications are brought out by the Reserve Bank of India.

Liabilities and Assets of Reserve Bank.

The weekly *Statement of Affairs of the Reserve Bank of India* gives the liabilities and assets separately of its Banking and Issue Departments, and is reproduced in several official and non-official publications, including newspapers. The Banking Department is concerned with the banking activities of the Bank; as such, its liabilities include capital, reserve fund, deposits by Governments, banks and others, bills payable, etc., and the assets include notes, coins, bills purchased and discounted, balances held abroad, loans and advances to Governments, scheduled banks and others, investments, etc. The Issue Department is concerned with currency issue; as such, its

liabilities include the total notes issued and assets cover gold coin and bullion, foreign securities, rupee coin and Government of India rupee securities. What is of significance to study in this part of the Statement is the ratio of gold coin and bullion and foreign securities to total notes issued. During 1950-51 this ratio was 56.31 percent and in 1951-52, 54.64 percent.

Liabilities and Assets of Scheduled Banks.

The weekly *Statement of Affairs of the Scheduled Banks* gives a consolidated account of the financial position of the Scheduled banks. To be a scheduled bank, a bank should be a company defined in Clause (2) of Sec. 2 of the Indian Companies Act, 1913, or a Corporation or Company incorporated by law outside India, and should have a paid-up capital and reserve of an aggregate value of Rs. 5 lakhs. Every scheduled bank has to maintain a reserve of at least 2% of its time liabilities and 5% of its demand liabilities with the Reserve Bank, in lieu of which it gets rediscounting facilities, and has also to furnish it with a weekly statement of its position. From these weekly statements the Reserve Bank prepares a consolidated *Statement,* which gives data on the following items:

1. Demand Liabilities in India;
2. Time Liabilities in India;
3. Cash in India;
4. Balances with the Reserve Bank in India;
5. Balances with other banks in India;
6. Money at call and short notice in India;
7. Investments in India;
8. Advances in India;
9. Inland bills purchased and discounted in India.

It will be seen that the total of items 1 and 2 constitutes the 'total liabilities of the scheduled banks' and the summation of items 6, 8 and 9 gives the 'total scheduled bank credit'. The chief interest in this Weekly Statement lies in studying the ratio of total bank credit to total liabilities. The following table summarises the position for the last few years:

Consolidated position of Scheduled Banks—Indian Union

(Lakhs of Rs.)

Average of Fridays	Total Liabilities	Total Scheduled Bank Credit	% of (3) to (2)	No of Reporting Banks
(1)	(2)	(3)	(4)	(5)
1948-49 ..	978,44	441,29	45.10	94
1949-50 ..	870,38	442,09	50.79	94
1950-51 ..	877,59	458,90	52.29	93
1951-52 ..	884,55	551,20	62.32	94

The summation of cash in hand and balances with the Reserve Bank (items 3+4) generally constitute 9.5% to 11.5% of the 'total liabilities of the scheduled banks'.

Liabilities and Assets of Non-Scheduled Banks.

The non-scheduled banks are not accounted for in the weekly *Statement of Affairs of the Scheduled Banks.* In terms of the Indian Companies (Amendment), Act, 1936, a banking company which is not a scheduled bank is required to transfer at least 20% of the declared profits to reserve, until it is equal to the value of paid-up capital, and is also required (if the company is a joint stock company) to maintain a cash reserve of at least $1\frac{1}{2}$% of its time liabilities and 5% of demand liabilities and to file a statement of the amount so held, with the Registrar of Joint Stock Companies. Liabilities and assets of the non-scheduled banks are also published on the same pattern as those of the scheduled banks, in the *Reserve Bank of India Bulletin.* In 1952-53, there were about 430 reporting non-scheduled banks. Their ratio of cash in hand *plus* balances with the Reserve Bank to their total liabilities varied between 7.5 to 8.5 percent during the year.

Other Banking Statistics.

The Trend and Progress of Banking in India is an annual report prepared by the Reserve Bank of India under Section 36(2) of the Banking Companies Act, 1949, reviews important developments in the field of banking during the year concerned and their impact on the national economy. It also discusses the changes in the structure of assets and liabilities of banks and the main features of the supervision and control exercised by the Reserve Bank over the banking system. It contains a statistical appendix, statements included in which are based on the various returns received periodically in terms of the Reserve Bank of India Act, 1934 and the Banking Companies Act, 1949, as well as on returns specially called for and other available information. The limitations of the data are explained in foot notes to the statements. Some of the important items on which these statements contain data are:

1. Liabilities and assets of the Reserve Bank;
2. Consolidated position of Scheduled Banks;
3. Liabilities and assets in India of Banking Companies;
4. Money rates in India;
5. Interest rates allowed by larger scheduled banks on deposits;
6. Analysis of investments of banks;
7. Analysis of advances of scheduled and non-scheduled banks;
8. Classification and distribution of banking companies;
9. Cheque clearances;
10. Velocity of circulation of deposit money.

The table below gives information on item 10 above:

Velocity of Circulation of Deposit Money

(Amount in crores of Rs.)

Year	Average demand liabilities of scheduled banks	Total cheque clearances	Ratio of (3) to (2)
(1)	(2)	(3)	(4)
1949 ..	609.7	6,325.4	10.4
1950 ..	594.3	6,278.2	10.6
1951 ..	604.1	7,878.0	13.0
1952 ..	556.9	6,853.2	12.3

Statistical Tables Relating to Banks in India contains comprehensive data on the working of Indian and foreign banks operating in India. It gives breakdown of data in respect of important balance-sheet items for each of the scheduled, non-scheduled and cooperative banks. Data on liabilities and assets of the Reserve Bank and the consolidated position of scheduled banks are repeated in this publication. It also gives information on the distribution of offices of various classes of banks.

Co-operative Banking Statistics.

Statistical Statements Relating to the Cooperative Movement in India is exclusively devoted to the statistical position of cooperative banking. It gives data, *inter alia,* on the number, membership, working capital and volume of business of State Cooperative Banks, Central banks, credit societies, land-mortgage banks and societies, and other types of cooperative societies. The *Indian Cooperative Review, Bombay Cooperative Quarterly, Madras and U.P. Cooperative Journals,* etc. also give relevant statistics.

The Reserve Bank is not empowered under the Act to finance agriculturists direct but it can provide the necessary finance to the cooperative movement through State Cooperative Banks. The Bank has also laid down the procedure which co-operative banks desirous of obtaining financial accommodation from it should fulfil. One of the conditions is that a State Co-operative Bank should maintain with the Reserve Bank cash balances the amount of which should not, at the close of business on any day, be less than $2\frac{1}{2}\%$ of the demand liabilities and 1% of the time liabilities of such a bank in India, and that it should submit to the Bank periodical statements prescribed for the purpose. These statements enable the Reserve Bank to present a consolidated picture of cooperative banking in the country in an annual *Review* and *Statistical Statements.*

The tables included in the *Statistical Statements* are divided into three parts. The General Statements give detailed statistics in respect of the various operations (such as number of members, loans advanced repaid and due, cash, investments, interest receivable, reserve fund, deposits, share and working capital, profit or loss, dividend paid, most usual rate of interest on borrowings and lendings, etc.) of different types of coopera-

tive institutions in the States; the Abstract Tables summarise the important data contained in the General Statement and, the Appendices give the data for years earlier than the one to which the main publication relates. The following table shows the number and membership of cooperative societies in India.

Number and Membership of Cooperative Societies
Indian Union

Type of Society	No. of Societies		No. of Members	
	1949-50	1950-51	1949-50	1950-51
CENTRAL				
Provincial and Central Banks and Banking Unions ..	512	520	2,08,340	2,28,006
Non-credit ..	2,113	2,236	13,47,102	14,73,203
Central Land Mortgage Banks ..	5	5	8,871	9,848
Provincial Unions or Institutes ..	18	19	29,235	29,195
Supervising Unions ..	496	518	25 231	27,887
AGRICULTURAL				
Credit	1,16,534	1,15,462	48,17,545	51,53,907
Non-credit ..	25,860	33,815	29,41,657	33,65,243
Primary Land Mortgage Banks ..	283	286	1,86,330	2,15,063
NON-AGRICULTURAL				
Credit ..	7,534	7,810	20,65,990	21 77,551
Non-credit ..	19,712	20,495	24,20,680	26,47,681
Insurance ..	27	23	1,28,814	1,55,575
TOTAL NUMBER OF MEMBERS OF PRIMARY SOCIETIES	1,73.094	1,81,189	1,25,61,016	1,37 15,020

The above table throws light on the structure of cooperative movement in India. It shows that while credit soc eties predominate among the agricultural societies it is the non-credit ones that dominate the non-agricultural field. It also shows the progress made by the cooperative movement in the

country in 1950-51 as compared with the previous year: both the number of societies and that of members increased in 1950-51. The extent to which these societies provide agricultural finance can be studied from the relevant tables of the *Statistical Statements*.

Rural Credit Survey.

In November 1951, the Reserve Bank of India sponsored an all-India survey extending over a period of eight months, with a view to collecting such data as would assist it, as well as the Central and the State Governments, to formulate an integrated rural credit policy for the whole of India. The survey was intended to cover, among other items, the following:

(i) Normal short, intermediate and long-term requirements of various classes of agriculturists;
(ii) Working of rural credit agencies, *viz.*, Government *taccavi* and other loans, cooperative credit organisations, indigenous bankers, etc;
(iii) Pattern of family savings and deficits; and
(iv) Various aspects of capital formation in the rural areas.

The enquiry was also intended to assess the effect of the enactments relating to agricultural debt relief, control of money lending, etc. on the rural credit system.

For the purpose of the survey, the whole country was divided into 33 homogeneous regions from the point of view of agricultural conditions. The selection of districts in each region was done on random sampling basis. In all 75 districts, situated in 20 States, were selected. From each district, 4 villages were selected with the help of random numbers, 2 from villages having no cooperative credit society and 2 from those having one. Thus, in all 600 villages were selected for the survey. In regard to demand for credit, the families in the selected villages were enumerated with the help of a general purpose schedule, and information was collected from them in regard to the size of cultivated holdings, family expenditure, capital investment in agriculture, borrowings, etc. On the basis of cultivated holdings, from each village 15 families were selected by random sampling method, detailed information from

whom was obtained through 7 schedules and 5 questionnaires. For studying the supply side of rural credit, information was obtained about finance by Government cooperative societies, land mortgage banks, indigenous bankers etc. to assess how far did these institutions adequately satisfy the credit requirements of agriculturists, what difficulties did they experience, and how should the shortcomings be remedied.

As regards the machinery to conduct the survey, one investigator conducted the actual survey of 4 villages, and one inspector supervised the work of 2 investigators. The inspectors worked under the direction of five Regional Controllers and four Deputy Regional Controllers. In some States, the Statistical Departments of the State Governments and in others some other private agencies (like the Gokhale Institute of Politics and Economics, Poona) also conducted the investigations, subject to general coordination by the Reserve Bank.

Currency and Bullion.

The annual *Report on Currency and Finance* is presented in two parts: Part I contains a resume of world economic development during the year under review and Part II a comprehensive survey of the general economic situation in India, followed by a series of chapters reviewing the trends and developments in domestic production and prices, money supply and banking, capital and bullion markets, public finance, foreign trade, balance of payments and exchange control. It gives a very large number of statistical statements on these subjects—the 1952-53 *Report* contained 90. Important among those relating to currency, finance, bullion and securities included the following:

1. Money Supply;
2. Absorption or return of currency and small coins;
3. Circulation of Notes;
4. Seasonal trends in currency circulation;
5. Liabilities and Assets of Reserve Bank of India;
6. Liabilities and Assets of Banking Companies;
7. Surveys of bank advances, deposits and investments;
8. Central, State and Railway budgets;
9. Public Debt;

10. Control of Capital Issues;
11. Security Prices and index numbers;
12. Bullion prices and imports and exports;
13. Balance of payments.

The table below gives the annual absorption or return of currency in the Indian Union during the four years ending 1951-52:

Annual Absorption or Return of Currency—Indian Union

(Lakhs of Rs.)

Year	Notes	Rupees	Small coins	Total
1948-49	(—) 7,84	(—) 4,31	(+) 24	(—) 11,91
1949-50	(—) 5,84	(+) 2,29	(—) 2,16	(—) 5,71
1950-51	(+) 83,89	(+) 5,82	(—) 3,20	(+) 86,51
1951-52	(—) 100,21	(—) 13,00	(—) 3,05	(—) 116,26

The weekly *Statement of Affairs* of the Reserve Bank of India relating to the Issue Department, dealt with earlier, also gives the currency position of the country.

NATIONAL INCOME

National income estimates measure the value of commodities and services turned out during a given period, counted without duplication. These estimates include all economic activity, whether it be the production of shoes or of aeroplanes or provision of justice. National income statistics provide a wide view of the country's entire economy as well as of the various groups in the population who participate as producers and income receivers. If available over a period of time, they reveal the basic changes in the country's economy in the past and suggest trends for the future. In the under-developed countries, like India, the problems with which the Government is chiefly concerned are starvation levels of output and gross

inequalities of income. The economic policy of such a country is designed to assure livelihood to its people, economic development and economic justice. For framing such policy, it is necessary to know the total amount of income generated in the country and how it is distributed as between different sectors. For imposing a retail sales tax, for instance, the Government has to estimate the magnitude of its possible collection. Now, if estimates of consumer's expenditure by various categories such as food, clothing, shelter, etc., are available, it should be possible to approximate what different rates of taxes, with different coverage, would yield. It should further be possible to evaluate the effects of such a tax on the disposition of consumer's incomes on the sales of taxed and untaxed categories.

Problems of National Income Estimation in India.

Problems of national income estimation in India are many. When calculating the value of output, one normally proceeds on the assumption that the bulk of the commodities and services produced in the country are exchanged for money. In India, however, a considerable portion of output is just bartered. The problem of imputation of value thus arises. The problem of measurability is further complicated in India by the fact that many producers have only a very remote notion of either the quantity or the value of their output. There is the general absence of the practice of keeping accounts among both the producers and consumers. An element of guess work, therefore, inevitably enters into the assessment of output, particularly in such sectors of the economy as are dominated by the small producer. Further, Indian economy lacks differentiation in economic functions inasmuch as a major portion of our economy consists of household enterprises performing functions which would normally fall under different industrial categories. For instance, many agricultural producers pursue other occupations in other industries, often in urban areas. Hence the customary classification of national income by industrial origin is different to make in India.

Apart from the conceptual difficulties relating to national income work mentioned above, there is a much larger problem of the non-availability of statistical data. Thus, prices and expenses data for agricultural and related activities are not

complete; data on factory establishments are available only for a part of the industries in which factory units are important; some of the items in the international balance of payments are rough estimates; no information is available on the structure of costs, on consumer expenditure, or on their savings; no useful data exist on distribution of income by size or on capital formation.

The above difficulties, both conceptual and statistical, in the way of national income estimation have been known for long. The Indian Economic Enquiry Committee of 1925 and the Bowley Robertson Committee of 1934 drew attention to them. The National Income Committee appointed by the Government of India in August 1949 also drew attention to them in their First Report published in April, 1951. The Committee made use of a number of expedients, assumptions and guesses, in order to overcome the many gaps in the statistitcal field. They had also the benefit of two pioneering studies on the national income of India made earlier by Dr. V. K. R. V. Rao. The estimates prepared by the Committee constituted, in the words of the Committee, "better estimates than any published so far" and this was said to be due partly to the availability of new material such as the Census of Manufactures and Marketing Reports and partly to the advance which had taken place in the whole field of national income work in the previous decade.

Methods of Measuring National Income.

Two methods of measuring national income are well known. The first is the Census of Production method, which involves an evaluation of the goods produced and services rendered. It consists in evaluating the net output of the various branches of productive enterprises—agriculture, forestry, fisheries, mining, industry, trade, transport, etc., at the point of production, adding the value of imports and subtracting that of exports. The value of home-produced goods and personal services, excise and customs duties, annual value of houses and the value of imports must be added, while the value of goods used for maintaining fixed capital or stocks of raw or finished goods intact, the value of exports, etc., must be deducted. The second method is known as the Census of Incomes method, which

involves the summation of individual incomes. In utilizing this method, all self-consumed produce, the annual value of houses, receipts in kind, etc., must be included; the incomes of those who pay income-tax and of those who do not are added together.* Both these methods require annual statistics of agriculture, manufactures, hand-made goods, mining, fishing, forestry, trade, transport, imports, exports, population, its occupational distribution, etc.

There is a third method also, that of Social Accounting, developed by Prof. Richard Stone. This method does not seek to build up a single total such as the National Income; instead, a study is first made of the classification of accounting entities, of the type of accounts which they keep and of the transactions that they make. With the help of this study, all the transacting entities of the economic system are classified into broad sectors, such as productive enterprises, financial intermediaries and final consumers. A series of accounts for each of these sectors is then set up, in which the separate entities represent economically distinct classes of transaction. Economic activity is represented by money flows and related book-keeping transactions between accounts. Lastly, the national income and similar other aggregates are obtained by selecting and combining the constituent entries in the account. This method is not feasible in India at present, since individuals as well as institutions in ths country seldom mantain accounts of transactions, unless, of course, compelled by law to do so.

Method adopted by the National Income Committee.

The broad pattern of the methods followed by the Committee had its basis in the methods adopted by Dr. V. K. R. V. Rao in his *An Essay on India's National Income: 1925-29* and *The National Income of British India: 1931-32*. Since the available data were not full and comprehensive, it was not possible to use either the Inventory or the Income method to cover the entire range of the economy. The Committee, therefore, estimated the total working force in 1948-49 and its distribution among different occupations. This occupational

* For details, please see *A Scheme for an Economic Census of India* by Dr. A. L. Bowley and Prof. D. H. Robertson, Pp. 9-12.

classification was made on the basis of the classification of the economy by industry. The term 'industry' included agriculture, services, and all other means of income-generation. The Inventory (or, the Value of Net Product) method was applied to estimation of national income from the following sectors:

(1) exploitation of animals and vegetation including animal husbandry and fishing;
(2) exploitation of minerals; and
(3) industry.

That is, national income arising from agriculture, mining, manufacturing and hand trades was estimated on the Inventory method. The Income method was applied to estimation of national income from commerce, transport, communication and other services. Broadly, 65% of the total national income was estimated on the Inventory method and the rest on the Income method.

National Income Estimates—1948-49.

The table below gives the estimates for 1948-49 framed by the Committee:—

National Income of the Indian Union by Industrial Origin: 1948-49.

Item	*Net output* (Rs. abja*)	*Percentage*
1. Agriculture	41.5	47.6
2. Mining, manufacturing and hand trades	15.0	17.2
3. Commerce, transport and communication	17.0	19.5
4. Other services	13.8	15.9
5. Net domestic product at factor cost	87.3	100.2
6. Net earned income from abroad	(—)0.2	(—)0.2
7. Net national output at factor cost=national income ..	87.1	100.0

* abja=100 crores.

Thus the estimated national income for the Indian Union for 1948-49 was Rs. 87.1 abja. The population in 1948 was taken as 341 millions on the basis of the estimates then made by the Registrar General. The *per capita* national income, therefore, worked out at Rs. 225. Although the population estimate was revised upwards subsequently, it did not affect the *per capita* national income estimate materially.

Agriculture (which includes agriculture proper, animal husbandry and ancillary activities such as processing, transport and marketing services performed by the cultivator, forestry and fishery) contributed roughly 48% or nearly half the national income. Commerce, transport and communications (which include posts, telegraph, telephone, railways, organised banking and insurance, and other commerce and transport) amounted to 19.5% or nearly a fifth of the total national income. When it is recalled that a large part of the commodity production does not enter into trade at all, being retained by the producer for his own consumption, the relevant importance of this sector would be clear. Commodity production, taken roughly as the total value contributed by agriculture, mining, manufacturing and hand trades, amounted to Rs. 56.5 abja or nearly 2/3rds of the national income. The Committee also estimated that out of the total national product of Rs. 87.1 abja, the consumer expenditure on food probably amounted to about Rs. 46 abja or nearly 58% of the national income. This figure was, however, given by the Committee for illustrating that the high production of consumer expenditure on food represented in a general way the undeveloped character of India's economy.

Issues for the Future.

All estimates framed by the Committee were obtained by calculating the value of each sector at its equivalent market value in terms of money. In actual fact, a large part of the commodities and services in India, as already said above, does not enter into the market at all; it is bartered and is, therefore, not subject to monetary transactions. In future estimates, attention may have to be given to two separate estimates for the monetized and non-monetized sectors. Again, the distribution of national income between urban and rural sectors is of

importance for policy purposes in India. This also is a problem to be studied in future. Lastly, the Committee found that it could not present the national income estimates in the form of social accounts as reliable information was not available on certain items like provision for depreciation, capital formation and savings in the private sector or consumers' current expenditure. It will be recalled that it is to fill such gaps in the economic and social statistics of India that the National Sample Survey organization has been set up.

Earlier Estimates of National Income.

Several estimates have been made, from time to time, of the national income of India. The first attempt was made on the basis of figures relating to the years 1867-1870 by Dadabhai Naoroji, who in his book on *Poverty and Unbritish Rule in India* estimated the national income of British India to be Rs. 20 a year per head. Subsequently, estimates were made by several Britishers until Wadia and Joshi worked out an estimate for 1913-14. They calculated the total income to be Rs. 1,087 crores and the *per capita* income to be Rs. 44.5. Thereafter, Shah and Khambhata and Findley Shirras made their estimates. Dr. V. K. R. V. Rao's calculations for 1931-32 then became available. According to him, the national income of British India for 1931-32 amounted to Rs. 1,688 crores with a margin of error of $\pm$ 6 per cent. The *per capita* income amounted to Rs. 62, with the same margin of error. If this estimate were inflated by the rise the wholesale price index number has since then recorded, it will be found that it would not be materially different from the National Income Committee's estimate of Rs. 255 per head for 1948-49. The conclusion is irresistible that between 1931-32 and 1948-49 the increase in national income of the country has been more or less absorbed by the increase in population.

The table below summarises the estimates of national income *per capita* framed in the past :—

Estimates of National Income Per Capita—India

Estimated by		Year for which estimated	National Income *per capita* in rupees.
Dadabhai Naoroji	..	1867-70	20
Earl Cromer & Sir David Barbour ..	..	1882	27
William Digby	..	1898-99	17.5
Lord Curzon	..	1900	30
Wadia & Joshi	..	1913-14	44.5
Shah & Khambata	..	1921-22	67
Findley Shirras	..	1922	116
V. K. R. V. Rao	..	1931-32	65
V. K. R. V. Rao*	..	1942-43	114

In studying the above table it should be remembered that the estimates are not comparable as between different years owing not only to differences in methods of calculation but also to the price levels obtaining in the different years. Also, the estimates do not all cover the same areas though they relate to India prior to Partition.

NATIONAL SAMPLE SURVEY

The Directorate of National Sample Survey (N.S.S.) was set up in 1950 under the Ministry of Finance. It has been set the task of collecting data on a national basis regarding production, consumption and other aspects of economic life, both in the rural and urban areas of India, and thus to make good the lacuna in the statistical information, so that planning and formulation of policies can be done in an intelligent manner. Since its inception, it completed six rounds of investigation by the middle of 1953, collecting various data bearing on the economic and social life of India. The results of the first round, which was completed in March 1951, were published in January, 1953.

The desire to collect information about the economic and social activities in a country as vast as India will remain a

* *vide*, the Commerce, dated the 26th February, 1944.

mere wish if collection of information for each one of the roughly 70 million households scattered over a territory exceeding a million and quarter square miles were the only possible method. Statistical theory and its application in concrete situations indicate the possibility of giving valid estimates for the country as a whole, or its smaller regions, on the basis of information collected by trained investigators from a comparatively small number of households selected at random according to the principles of statistical science. The basic approach of the N.S.S. is, therefore, one of making enquiries from the randomly selected households, in the randomly selected villages. More than 300 trained investigators scattered all over India collect the information under the supervision of the N.S.S. The technical work including the design of the survey, the scrutiny, processing and analysis of data is done at the Indian Statistical Institute, Calcutta. The investigators collect information on several aspects of urban and rural economy by visiting the sample households. They interview the members of the households to elicit the required information and after checking and cross-checking the information given to them in order to test its reliability, fill up a number of carefully designed schedules. This method of collecting information by interview is used in the Population Census and has been used in some of the other important all-India surveys such as the Agricultural Labour Enquiry and the Rural Credit Survey.

The successive rounds of the N.S.S. are designed with a view to supplying, from year to year, repeated or seasonal estimates of important economic factors to indicate economic trends. The N.S.S. also proposes to conduct from time to time special surveys relating to particular subjects in collaboration with the concerned Central Ministries. Among the various special or *ad hoc* surveys in which the N.S.S. has been engaged may be mentioned the following: Survey of displaced persons in West Bengal and Bombay (Ministry of Rehabilitation); collection of information on country-wide basis for the Press Commission (Ministry of Information and Broadcasting); survey of unemployment in Calcutta and other urban areas (Planning Commission); survey of housing (Ministry of Works, Housing and Supply); some preliminary investigations

in connection with the proposed land census (Planning Commission); work in connection with the Mysore Population Study (U.N. and Ministry of Health).

In the Second Round of the survey, which was in operation from April, 1951 to June, 1951, data have been collected on the pattern of consumption and distribution of rural households by their consumer expenditure, the extent of monetization of consumption flows, estimates of rural capital formation, and the distribution of households by the size of land operated or managed by them.

The Sixth Round of the survey which commenced in May, 1953 sought to collect comprehensive information on a wide range of subjects: consumer expenditure; other household expenditure; possessions including lands, tanks, agricultural implements, live-stock, poultry, etc.; fertility; birth, deaths and diseases; manufacturing establishments; agriculture and animal husbandry; small scale industry and handicrafts; transport; trade; professions and services, etc.

The First Round of the Survey.

The results of the First Round of the N.S.S. were published in January, 1953 through the *General Report No. 1 on the First Round, October, 1950—March, 1951.* It contained 96 tables bearing *inter alia* on consumer expenditure per household, value of consumption of each article of consumption, expenditure on different items of consumption, consumption of foodgrains, salt and milk, value of agricultural crops and live-stock products, costs, gross income from different services, and area under foodgrains.

One finding of the First Round which received considerable attention was that the total consumption of foodgrains in India out of home-grown produce was 60 million tons during the year July, 1949, to June, 1950. As compared to this figure, the official estimate of production of foodgrains for that year was stated to be an under-estimate to the extent of 20% to 25%; but in this connection the report added: "This result may be purely fortuitous; or due to some bias in collecting the information. A final verdict can not be given immediately on the basis

of the first round alone".† In the same connection it said elsewhere, "......this does *not* mean that there is a surplus of (say, about 12 or 16 million tons of) foodgrains in the country."* The *Report* did not give any idea of the margin of error in the N.S.S. estimate of consumption, so that nothing can be said about the significance of the estimate itself. Another factor which put a limitation on the results of the consumption inquiry was that the selected households were asked about the size of their consumption only after a few months of the end of the consumption year: it was during October, 1950, and March, 1951 that the households were asked what they consumed during the year July, 1949 to June, 1950. It is possible that the households did not remember their actual offtake over such a long period as a year, particularly when the consumption year had been over a few months ago. It is interesting to note that in the Second Round of the N.S.S. the reference period was changed from one year to one week for items of general consumption "to see if the householders were able to remember their consumption more correctly for this short period."@ To summarise with the Report "......it was not clear whether the figures of consumption of cereals and cereal products were correct or under-estimates."@

The main findings of the First Round may now be summarised:

(1) On an average, each household in the rural areas comprised 5.21 people, living in the household for a month or more, of whom a little over $\frac{1}{4}$th were earners, about 1/6th earning dependents and more than half non-earning dependents.

(2) The consumer expenditure in rural areas for 1949-50 worked out at Rs. 220 per person per year. The level of average consumer expenditure per person varied considerably as between regions, the highest being Rs. 314 in North-West India (Rajasthan, Pepsu, Ajmer, Delhi, Punjab, Bilaspur and Himachal Pradesh) and the lowest being Rs. 198 in Central India (Madhya Bharat, Madhya Pradesh, Hyderabad, Bhopal and Vindhya Pradesh).

† *Cf.* p. 42 of the *General Report No. 1.*
* *Cf.* p. 32 of the *General Report No. 1.*
@ *Cf.* para 9.3, page 81 of *General Report No. 1.*

(3) As much as 2/3rds of the household expenditure went to food, roughly 1/10th to clothing and the remaining $\frac{1}{4}$th accounted for all other items of expenditure. Ceremonials absorbed a little over 7% of the total expenditure.

(4) The consumption of milk per person in the country as a whole amounted to less than 2 seers per month. It was the highest, about 66 seers per person per year in North-West India, and the lowest, about 12 or 13 seers in South India (Madras, Mysore, Travancore-Cochin and Coorg) and East India (Bihar, Orissa, West Bengal, Assam, Manipur, Tripura and Andaman and Nicobar Is.). Likewise, the expenditure on milk and milk products per person per year was nearly Rs. 60 in North-West India, while it was as low as Rs. 9 in South India and Rs. 7 in East India.

(5) The variations in expenditure on tobacco and salt as between different zones were small as compared to those in the case of milk.

(6) The average expenditure on clothing was about Rs. 21 for the country as a whole, the highest being Rs. 31 in West India (Bombay, Saurashtra and Kutch) and the lowest being Rs. 15.5 in East India.

(7) The average rate of consumption of foodgrains in the rural area worked out at 26.4 maunds (2,163 lbs.) per household per year, that is about 18.3 or. per person per day. As against this, the estimated production of foodgrains as deduced from the replies received from farmers worked out at a little less than 20 maunds per household. The report held the latter estimate as a serious under-estimate.

(8) The inquiry related to rural household alone and to arrive at the consumption of the country as a whole the *per capita* urban consumption was taken to be 76% of the rural consumption per person. The all-India consumption of foodgrains thus averaged to 17.6 oz. *per capita* per day.

Design of the Survey—First Round of N.S.S.

As the material available was different for different regions, different types of "frame" were adopted; the methods of selecting the sample-units in different parts of the country

were, therefore, different. The probability of being included in the sample thus differed from region to region which made the design of the survey somewhat complicated and less efficient in the First Round. Varying sampling fractions were used for selecting the villages.

The sample units were selected in two stages. First, the villages were selected after suitable stratification. Within each sample village, all or a sub-sample, whichever was less, were stratified into agricultural and non-agricultural classes. Then, sample households were selected for each of these classes.

Altogether 1,833 sample villages were selected for the whole of India. In each stratum the number of sample villages was made a multiple of 3, so that they could be broken up in the ratio of 2:1. Thus out of the 1,833 sample villages, two groups were formed, one of 1,189 villages, and the other of 644 villages. The first group was allotted to the N.S.S. and the second to the Gokhale Institute of Politics and Economics, Poona. The N.S.S. and the Poona Institute conducted the inquiry in their respective groups of sample villages with their separate schedules and separate staff. Out of the sample villages allocated to the N.S.S. actually 1,111 were surveyed. The *General Report No. 1* relates to the N.S.S. group of villages.

The different phases of sampling of households in the N.S.S. villages were as follows:—

(1) All the households in a sample village were listed;

(2) A random sample of 80 households was selected for collecting information relating to occupation;

(3) These 80 households were then classified into two groups: agricultural and non-agricultural;

(4) Out of the agricultural households, 8 were selected at random. Similarly, 8 were selected from non-agricultural households. Household schedule for general particulars was then completed from each of these 16 households.

(5) Two of these 8 agricultural households and 3 of the 8 non-agricultural households were then selected at random for completing the schedules for household enterprises.

(6) Of the remaining 6 agricultural households, one was selected; and of the remaining 5 non-agricultural, two were

selected, that is, three in all for completing the schedules for detailed information on consumer expenditure.

Poona Schedules.

It has been said above that a group of 644 villages was allocated to the Gokhale Institute for conducting the national sample survey. The Institute tried its own schedules in these villages, and the results of the survey are contained in the *Report on the Poona Schedules of the National Sample Survey (1950-51)*, brought out by the Institute. The tables given in the *Poona Report* are based on only one visit to the selected villages some time from October, 1950 to March, 1951. The Report adds: "Most items of the Poona Schedules refer either to the day of visit or some short period immediately preceding that day; and many items in them are liable to seasonal variations. It is for this reason that we have purposefully presented these tables as relating to only one month (four weeks) or a week or only three days as the case may be and have not inflated them (presumably, unlike the *General Report No. 1* referred to earlier) to be brought to an annual basis." (p. 160 of the *Poona Report*). The tables are based on a survey of 585 villages, and give information on a number of items, including age distribution, occupational distribution, livestock and implements, employment, production, sales, cash receipts and expenditure, and household expenditure.

In a note, reproduced in the *Poona Report* (*vide* pp. 204-205), Prof. Gadgil said:

"1. We are not getting or are not likely to get in the immediate future, through the National Sample Survey, as at present conducted, information which is reliable enough for the purpose for which it is intended to be used.

"2. The main reasons for this are:—

(*a*) The agency employed is not familiar enough with rural conditions;

(*b*) the programme of work, even in the Poona Schedule, does not enable it to establish contacts in the proper way with the respondents;

(*c*) adequate provision has not been made for training and supervision.

"3. ....too many aspects have been considered in the survey........

"4. Therefore, if Government wants to maintain activity in the field and incur expenditure, the whole plan should best be thought out afresh........"

The Poona Schedules were expected to provide a check on the findings of the main N.S.S. enquiry into 1,111 villages. It was also expected that a comparison of the two results would enable a better appreciation of the right method of obtaining statistical data from a vast multitude of India's population. The *Poona Report* falsified these expectations. It gave a discourse on methodology and showed how it was defective for the purpose for which it was employed. It gave some data too, but it is difficult to institute any valid comparison between these data and the data given in the *General Report No. 1.*

GENERAL

From the foregoing description of the nature and content of official statistics in India, the chief sources that one should consult for the statistics in which he is interested must have become known. It should be remembered, however, that statistics relating to a particular sector of India's economy can be found in more than one publication, that there are certain all-purpose publications which contain data on several economic sectors and that the Government are trying to meet the public demand for quick supply of data through weekly and monthly bulletins or journals, and wherever possible, important data are being put out through the daily press and the radio. The names of a few of such general publications are given below, and one would do well to familiarize himself with their contents:—

I. Publications of Central Statistical Organisation—
   1. *Statistical Abstract of India* (annual),
   2. *Monthly Abstract of Statistics,*
   3. *Weekly Bulletin of Statistics.*

II. Publications of Reserve Bank of India—
   1. *Reserve Bank of India Bulletin* (monthly),
   2. *Report on Currency and Finance* (annual).

III. Publications of Ministry of Commerce and Industry—
   1. *Journal of Commerce and Industry* (monthly),
   2. *Indian Trade Journal.*

IV. Publication of Directorate of Economics and Statistics—
   1. *Agricultural Situation in India* (monthly).

V. Publication of Labour Bureau—
   1. *Indian Labour Gazette.*

CRITICAL APPRAISAL OF INDIAN STATISTICS

Official statistics in India have been criticised on several counts. These criticisms may be said to centre round the following:—

1. Their accuracy is questionable.
2. They are inadequate.
3. They are inconsistent and incomplete.
4. There is lack of co-ordination in their collection and presentation.
5. The data published are not self-explanatory.
6. There is delay in their publication.

Accuracy of Indian Statistics is often questioned on the ground that the primary agency employed for the collection of some of the data is hardly trustworthy. This is said to be the case with agricultural statistics, which have, for long, been collected by the revenue staff in the temporarily settled areas and police staff in the permanently settled areas. The position in regard to reliability of agency has, however, considerably improved in recent years: Scientifically trained statistical staff now collects these statistics in a large number of States by employing objective methods and enforcing more strict supervision; in the premanently settled areas of Bihar and West Bengal they are no longer being collected through the village *chaukidar*. In the States where there is still room for improvement, as in Orissa, Saurashtra, Rajasthan, etc., steps in the right direction are already under way. Likewise, statistics of market prices of various commodities, which were formerly being collected by the revenue agency in the States, are now being collected in most cases by the Marketing, Civil Supplies,

Procurement or Economic Intelligence staff, with adequate arrangements for their checking and supervision. Similarly, in several other spheres of economic statistics, determined efforts are being made to improve the reliability of data by appointing scientifically trained staff. That there is still some leeway to be made up, such as in the collection of vital statistics, is, no doubt, true.

The other ground on which the accuracy of official statistics is often questioned is the impression that scientific methods are very rarely applied to the analysis of primary data so that conflicting conclusions are often drawn from the same set of figures at different times or by different interpreters. This impression, in fact, arises from a sort of belief that official statistics are motivated or biased, that they are so presented as to glorify the working of the department to which they relate. Such belief is, however, losing ground with the entry of trained statistical personnel into the different statistical organisations to collect, analyse and interpret statistics, with greater coordination between these organisations and with the increasing association of non-official organisations.

Economic statistics in India have been criticised for their being inadequate. The Indian Economic Enquiry Committee, 1925, examined the material then available for estimating the economic condition of various classes of people in India and found that general statistics other than those relating to production were more or less complete, that statistics of production were satisfactory in some respects but incomplete or totally wanting in others, and that no satisfactory attempt had been made to collect the statistics in income, wealth, etc. The Bowley Robertson Committee of 1934 remarked that though in some branches of statistics determined efforts were being made to improve the accuracy and scope of information, in others they were unnecessarily diffuse, incomplete or misleading, while in important fields general information was almost completely absent. The Commission which enquired into the Bengal Famine of 1943 also had to complain against inadequacy of data. The Report of the National Income Committee, 1948-49, pointed out the various lacunae in Indian statistics that they came across, and following the Report, a National Sample Survey Organisation came into being to pro-

vide a continuing stream of data on different sectors of Indian life. When the question of arresting the growing unemployment in the country came up before the Government in the middle of 1953, it was found that no adequate statistics on the subject existed, and the Government had to ask the National Sample Survey Organisation to conduct special surveys. The main hurdle in the way of achieving the desired adequacy of national statistics is, no doubt, financial; but as the country is advancing towards a Welfare State and as the realisation is growing that information must precede economic planning, the scope of Indian statistics is gradually expanding.

Incompleteness and inconsistency constitute another shortcoming of economic statistics in India: Incompleteness in space has considerably diminished since Independence, thanks to the merger and integration of the former Indian States. Incompleteness in regard to scope or items of information is more or less the same thing as inadequacy and has been dealt with in the previous paragraph already. As regards inconsistency, it will now be the job of the co-ordinating machinery at the Centre and in the States to provide the necessary remedy. Of course, during the transitional period when uniform standards and definitions are still being built up, nature, scope and coverage of data are being widened and methods of their collection are being improved upon, it is natural that continuity or consistency with the past should break. All the same, continuity of statistics is as important as their reliability and adequacy, and whenever a break with the past becomes necessary, it will be useful to continue both the old and new series for some time so as to give the reader an idea of the magnitude of difference between the two sets of figures.

Co-ordination in Indian Statistics was not hitherto possible in the absence of a coordinating agency. This deficiency has now been made up. The extent to which a coordinated picture of Indian statistics would be presented in future would naturally depend on the success attending the work of the co-ordinating authority.

Official statistics were not found to be self-explanatory in the past and their exact scope, methods of compilation and limitations were not widely known. To make up this deficiency, the Office of the Economic Adviser under the Ministry of Commerce

and Industry published three *Guides to Current Official Statistics* relating to production, trade, transport, public finance, etc. The last of these Guides was published as far back as 1948. The Guides, therefore, need to be brought up to date. The Directorate of Economics and Statistics under the Ministry of Food and Agriculture have also issued Guides to certain commodity statistics like cotton, jute, sugarcane and oilseeds, and when statistics relating to distribution of foodgrains in the country were published by this Directorate for the first time in the *Indian Food Statistics, 1949*, a detailed account of what the data included in each table of the book signified, was also given. When the Labour Bureau published the all-India average working-class cost of living index number towards the end of 1952 and the index of earnings of factory workers in India in February, 1953, it also published an account of the scope and method of construction of these index numbers. In fact, most statistical publications are now giving explanations regarding the nature, limitations, etc., of the different data either by way of footnotes or in explanatory or introductory notes or in appendices.

The last but not the least important criticism of official statistics in India is the delay in their publication. Figures have, in many cases, been found to be completely out of date by the time of their publication. If they are not meant to be purely academic, they should be recent when published. Delays in printing during World War II were understandable. But, that delays should continue even long after the War is painful. The demand for uptodate statistics is now growing and the Government seem to be resorting to certain devices of publishing, at any rate the important statistics, as quickly as possible. To this end, some statistical organisations have started issuing monthly journals wherein are published current and important statistical data; some are even issuing their bulletins in cyclostyled form to save the time taken by the printing press; and, in several cases data are being regularly issued through the agency of the press and the radio. Some delay in the publication of data whose compilation is dependent on the cooperation of States or districts, is, no doubt, unavoidable. Delay in the receipt of data by the Central Government from even one, or, at any rate, an important, State, would hold up the publication of the data

for the rest of the country. This applies equally well to a State publication which depends on receipt of data from all its districts. To quench the people's thirst for statistics, the system needs to be improved so that delays on the part of reporting agency and the printing press are minimised, if not done away with altogether.

It will be seen from the above that the present position of Indian statistics is not so bad as it used to be in the recent past and that the Government are attempting to make them truly national. That some improvement has taken place in recent years is true; but that there is still room for improvement cannot be denied. Improvement always takes time; there is no easy and quick road to that destination. The hopeful sign is that there is determination to achieve the objective.

EXERCISES.

(1) Give an account of the system for the collection, compilation and publication of statistics in India, and discuss in this connection the merits and demerits of a decentralized system.

(2) Attempt a lucid note on the statistical organisation in India pointing out how the needs of co-ordination in statistical work are met at the Centre and in the States.

(3) Write a brief critical note on the 1951 census of population. (B. Com., Alld., 1952).

(4) Discuss the possible value of Census Reports to producers, manufacturers and businessmen. How can the Indian Census Reports be made more useful to these people?
(M. Com., Alld., 1943).
(B. Com., Nagpur, 1945).

(5) Give a specimen of the enumeration slip used in the 1951 census, and explain the meaning and scope of each question therein.

(6) Discuss the main features of population statistics in India. What suggestions would you offer to make them more reliable and useful. (M.A., Alld., 1951).

(7) 'The Phoenix system is in fact a financial mistake as well as an intellectual crime'.

(Census Commissioner, 1941).

How far do you agree with the above criticism of the Indian census? In what respects is the system of conducting the 1951 census an improvement over the 1941 census?

(M.A., Pb., Sept., 1951).

(8) What statistical data are necessary for calculation of the net reproduction rate? What is the deficiency in the existing Indian data in this respect?

(M.A., Alld., 1951).

(9) How are vital statistics collected and compiled in India? What are their defects? Give suggestions for their improvement.

(10) Emphasize the role which vital statistics play in the estimation of annual changes in population in a country. Can the vital statistics available in India be utilised for that purpose? How would you put them on an even keel?

(11) 'The available agricultural statistics in India are incomplete and inaccurate in so far as (*a*) data relating to acreage and production in non-reporting areas are wanting, (*b*) information in respect of permanently settled areas is far from satisfactory, and (*c*) the level of accuracy of output figures leaves much to be desired'.

Comment on this statement, suggesting measures for effecting improvements in each of these directions.

(B. Com., Alld., 1949 & 1951).

(12) The traditional method of estimation of yield of crops in India is said to suffer from two defects, *viz.,* it is subjective and the margin of error in the resulting estimate is not known. What has been done to remove these defects of yield estimation in India in recent years?

(13) Give a brief account of the main improvements that have recently been effected in agricultural statistics in India. Along what lines would you like these statistics to be further improved?

(14) Define a normal yield and describe the official method of determining it. What do you consider to be the defects of the method and how would you remove them?

(M.A., Raj., 1950).

(15) What do you understand by the term 'Indian Agricultural Statistics'? Outline their shortcomings and give concrete suggestions to remedy them.

(M.A., Raj., 1951).

(16) Why are agricultural statistics in the temporarily settled areas in India said to be comparatively more reliable than those in the permanently settled areas?

(M.A., Pb., April, 1952).

(17) Write a lucid note on either the system of crop forecasting in India or the adequacy and reliability of data available on agricultural prices and wages in India.

(M.A., Pb., Sept., 1952).

(18) What important statistics of food production are available in India. How are they compiled, and in what official publications are they found?

(M. Com., Alld., 1952).

(19) How are the statistics relating to livestock and poultry collected in India? When was the last enumeration of these statistics made? What is the size of our livestock wealth according to this enumeration?

(20) What is meant by Census of Production? Why is such census taken? How far is the Industrial Statistics Act adequate from the point of view of holding this census in India?

(M. Com., Alld., 1947).

(21) Write a lucid note on the nature and scope of industrial statistics in India.

(B. Com., Alld., 1953).

(22) The Census of Manufactures in India is said to show the way but not to complete the journey yet.—Explain the significance of this statement.

(23) What is meant by the 'value added by manufacturers'? How is it estimated in India?

(24) Explain the importance of wages statistics and examine what information is available in India on industrial wages.

(25) Agricultural wages are now said to be available on a uniform basis for skilled and unskilled workers in different States. Name the chief sources you will consult for getting these data.

(26) What was the object of conducting the Agricultural Labour Enquiry in India? How was the Enquiry organised and conducted?

(27) Explain the importance of 'Price Statistics', and examine the nature and scope of the data relating to them available in India.

(M. Com., Alld., 1947).

(28) Great advance is said to have been made in recent years in the matter of collection and dissemination of prices data in India. Write a lucid note giving an account of this advance and giving suggestions for further improvement.

(29) What are the chief sources for data relating to different categories of prices in India? Point out their limitations.

(30) What are the chief sources for the overseas and inland trade of India? What type of information is available in them?

(31) The foreign trade statistics of India are said to be one of the oldest, best organised and most comprehensive statistics in India. Do you agree?

(32) Examine the nature and scope of statistics of inland trade available in India.

(33) Which publication would you consult to find (*i*) the number of cooperative and land mortgage banks in different Indian States, (*ii*) the rail-borne inland trade for the U.P., (*iii*) the annual absorption of small coins in India, and (*iv*) the wheat area irrigated in the U.P. and Bihar?

(M.A., Alld., 1952).

(34) What statistics are necessary to keep under review the purchasing power of a country's currency. Write a critical account of such statistics in India.

(M.A., Agra, 1945).

(35) Examine the nature and scope of financial statistics available in India.

(36) Statistics relating to Public Finance in India are said to be quite comprehensive and detailed.—How far do you agree with this statement? What type of data are available on the subject?

(37) Where will you find the data in respect of banking and currency in India? Examine their nature and scope and suggest measures for their improvement.

(38) What are the methods usually adopted for measuring the national income of a country? Which of these, in your opinion, is suited to Indian conditions?

(M.A., Alld., 1950).

(39) What is national income? What statistical methods of its estimation are known to you? Give a lucid account of the method actually adopted in any of the recent official or non-official estimates framed for India.

(M.A., Pb., Apr., 1952).

(40) What are the special problems of national income estimation in India? Describe briefly the various methods followed for the calculation of Indian income.

(M. Com., Alld., 1952).

(41) Describe briefly the method followed by the National Income Committee for framing estimates of national income in India for 1948-49. How far does this method differ from the one recommended by the Bowley-Robertson Committee?

(B. Com., Alld., 1952).

(42) "The methods followed in the estimation of India's income have been necessarily governed by the availability of data. It has, therefore, not been possible to use either the 'inventory' or the 'income' method to cover the entire range of economy."—*First Report of the National Income Committee, April, 1951.*

In view of the above observations of the Committee, explain what procedure it adopted to frame the estimate of the national income of India in respect of the different sectors of economy for 1948-49.

(B. Com., Delhi, 1953).

(43) What is National Sample Survey? What type of surveys have been conducted or undertaken by this body?

(44) Give an account of the design of the survey adopted for the First Round of the National Sample Survey. What were its main findings?

(45) Name some general publications in which current economic indicators could be found, indicating the nature of data available in them.

(46) Give a critical appraisal of the official statistics in India, mentioning also the steps being taken to remove their shortcomings.

(47) Describe briefly the nature of data available in the following:—

1. *Indian Agricultural Statistics.*
2. *Accounts relating to the Foreign (Sea, air and land) Trade and Navigation of India.*
3. *Monthly Statistics of the Production of Selected Industries in India.*

(48) What are the principal sources of statistical data for India? Examine, suggesting improvements, the materials available under the following heads: (*a*) Agriculture, (*b*) Wages and (*c*) Prices.

(M. Com., Alld., 1943).

(49) What statistical information is available in India with regard to (*a*) Currency and finance (*b*) population of livestock and (*c*) consumption of foodgrains.

(50) Describe the objects of and the methods adopted for conducting any two of the following surveys in India:

(1) Random Sample Crop-Cutting Surveys,
(2) Agricultural Labour Inquiry,
(3) Rural Credit Survey,
(4) National Sample Survey.

(51) In the case of one of the following censuses, explain how the recent census was organised:

(1) Population Census,
(2) Livestock Census,
(3) Census of Manufactures.

CHAPTER XXIII.

INDIAN INDEX NUMBERS

Index numbers constitute a very useful device of studying trends of different phenomena over a period of time. The construction of indices in India dates back to the last quarter of the 19th century. With the year 1873 as base, index numbers of (*a*) Prices of Exported and Imported Articles, (*b*) Retail Prices of Foodgrains and (*c*) Wholesale Prices used to be compiled and published in the *Index Numbers of Indian Prices*, a quinquennial publication with annual supplements. During World War I, Calcutta and Bombay Wholesale Price Index Numbers came into being with July, 1914 as the base. After World War I, it was the turn of the cost of living index numbers which began with the Bombay Working Class Cost of Living Index Number. During and after World War II, the scope of index numbers got extended to several new fields including production, profits and foreign trade. For wholesale prices, the Economic Adviser first issued the sensitive index based on 23 articles, and, with effect from 1947, replaced it by a general purpose index. Between December, 1952 and August, 1953, two new series and one revised series of index numbers became available: the new series included the All-India Average Working Class Cost of Living Index Numbers and the Index Numbers of Earnings of Factory Workers both compiled by the Labour Bureau; the revised series related to Index Numbers of Security Prices compiled by the Reserve Bank of India.

Thus the volume of index numbers has been growing in India; the growth can be easily said to be phenomenal since the World War II. It is intended to describe here the nature, scope and method of construction of the more important series. The series chosen are also the current ones, except in one or two cases. A few non-official series have also been included. An attempt has also been made to provide tables showing the relevant index numbers for some period of time, to enable familiarization with the finished product. These tables could

be only brief; the details can always be seen in the sources cited. These tables generally give annual indices even where the frequency of compilation of an index is much less, say a week. It should be remembered that index numbers compiled over small frequency can be averaged to give indices over large frequency; a weekly index series can, for instance, be averaged to yield a monthly, quarterly and annual series. Annual series have been given here as they enable comparison over longer period of time than would be possible through series of less frequency.

The index numbers chosen for discussion are the following:—

I. AGRICULTURAL PRODUCTION—
1. Ministry of Food and Agriculture series,
2. Reserve Bank of India series,
3. The *Eastern Economist* series,
4. F. A. O. series.

II. INDUSTRIAL PRODUCTION—
1. Ministry of Commerce and Industry series,
2. The *Eastern Economist* series.

III. INDUSTRIAL ACTIVITY—
1. The *Capital* series.

IV. COMMODITY PRICES—
1. Ministry of Food and Agriculture Harvest Prices series,
2. Economic Adviser's Wholesale Prices series,
3. Calcutta Wholesale Prices series,
4. Labour Bureau Retail Prices (Urban Centres) series,
5. Labour Bureau Retail Prices (Rural Centres) series.

V. SECURITY PRICES—
1. Reserve Bank's old series,
2. Reserve Bank's revised series.

VI. WORKING CLASS COST OF LIVING—
1. Labour Bureau series for 15 centres,
2. Labour Bureau series for Plantation Workers,
3. State Governments' series,
4. Labour Bureau All-India series.

VII. EARNINGS OF FACTORY WORKERS—

1. Labour Bureau series.

VIII. INDUSTRIAL PROFITS—

1. Ministry of Finance series.

IX. FOREIGN TRADE—

1. Department of Commercial Intelligence and Statistics series,
2. Reserve Bank of India series.

I. AGRICULTURAL PRODUCTION

The following series of index numbers of agricultural production in India are available:

(1) Ministry of Food and Agriculture series,
(2) Reserve Bank of India series,
(3) The *Eastern Economist* series,
(4) F. A. O. series.

All these index number series relate to gross production of crops. The production figures considered are those published in the *Crop Estimates* or the *Estimates of Area and Production of Principal Crops in India* issued by the Directorate of Economics and Statistics under the Ministry of Food and Agriculture. In all the series, individual commodities are grouped into certain groups and separate figures are published for the individual commodity indices, group indices and the general index. All the series are naturally compiled on annual basis. The index numbers for the latest year, being based on Final Estimates, have to be provisional and subject to revision next year when Revised Estimates become available.

The formula used in these series is generally of Laspeyre's type for quantity index, where base-year weights are employed (*Cf.* p. 246, Chapter XI). The manner of determining the weights is almost similar in the first three series, but different in the F.A.O. series; foodgrains are included in all the series, while the coverage of non-foodgrains slightly varies and so does the number of groups into which non-foodgrains are classified; the base period is the average of pre-War years in all cases, though the actual years comprising the base period are not identical. The Ministry's series when first published in

January, 1949, gave index numbers for the then Provinces and Indian States separately, besides the all-India index. The other series are all-India series only.

Ministry of Food & Agriculture Index.

This series is available from 1939-40 to 1945-46, in the *Agricultural Situation in India,* January, 1949 issued by the Directorate of Economics and Statistics. The series has since been suspended, pending the revision of the scheme for its construction. The base period was the average of the quinquennium 1934-35 to 1938-39.

This series was based on 19 principal agricultural commodities, arranged in two major groups: Foodgrains and Non-Foodgrains. The former group included 8 major foodgrains, *viz.,* rice, jowar, bajra, maize, ragi, wheat, barley and gram; the latter group included 5 oilseeds (sesamum, groundnut, rape and mustard, linseed, castorseed), 2 fibres (cotton and jute), 2 beverages (tea and coffee) and 2 miscellaneous crops (sugarcane and tobacco). The production figures were all for the Indian Union, suitable adjustments having been made for the figures in the pre-Partition years. The total production of each of these commodities was expressed as a percentage of the average production of the commodity during the base period. In all cases, weighted geometric mean of the production relatives was taken to represent the index number of production. The weights assigned to different commodities were proportional to the total value of each commodity during the base period. In evaluating the average value of production, the average harvest price during the base period was used. Where such price was not available in respect of a particular commodity, it was estimated from the wholesale price. The general index was obtained by combining the two group indices. The proportion of weights assigned to the Foodgrains group and the Non-Foodgrains group in the all-India series was as 2:1. In the former group, about half the total weightage went to rice alone.

Reserve Bank of India Index.

This series is available from 1939-40 to 1946-47 for India prior to Partition with the average of the triennium 1936-37 to

1938-39 as the base, from 1947-48 onwards for the Indian Union with the same base, and from 1949-50 onwards for the Indian Union with the year 1948-49 as the base. This series is published in the *Reserve Bank of India Bulletin,* generally in its December issue.

The series is based on 17 commodities, divided into 5 groups: Foodgrains, Beverages, Oilseeds, Fibres and Others. The commodities included are the same as in the Ministry of Food and Agriculture series, except that sugarcane and tobacco are excluded from that list and rubber included instead. The 8 foodgrains appear as 7 in this series, jowar and bajra having been combined under millets. The index is a weighted index, value of crop production constituting the weights. The weights are as under:

Weights in Reserve Bank's Agr. Production Index.

Commodity	Weight	Commodity	Weight
Rice	38	Sesamum	1
Jowar & bajra	12	Groundnut	7
Maize	2	Rape & Mustard	2
Ragi	2	Linseed	1
Wheat	14	Castorseed	0.3
Barley	4	Cotton	3
Gram	7	Jute	2
		Tea	4
		Coffee	0.4
		Rubber	0.1

Foodgrains, in this series, get a total weightage of 79 out of 100 of which nearly half goes to rice alone.

The Eastern Economist Index.

The *Eastern Economist* is a weekly, non-official, economic journal. The series compiled and published by this journal is also available from 1939-40, with the same base period as that of the Reserve Bank series. It was published in the Special Budget Number, 1952-53, of the journal, again in similar Number for 1953-54 and then in the issue for week ended the 31st July, 1953.

Besides the Foodgrains group, this series contains three other groups—Fibres, Oilseeds and Miscellaneous—which include all the commodities covered by the Non-Foodgrains group of the Ministry's series except castorseed. The *Eastern Economist* index is also a weighted index, the weights being proportional to the values of commodities during the base period. The value of each commodity is expressed as a percentage of the total value of all the commodities included in the index.

The F.A.O. Index.

The Food and Agriculture Organisation of the United Nations also constructs these index numbers for India, among other countries of the world, with the average of the years 1934 to 1938 as base, and publishes them in the *F.A.O. Year Book.* These indices relate to gross production of crops but the number of commodities included is large. Commodities are grouped under these heads: Grains, Starchy roots, Sugar, Pulses and oilcrops, Nuts, Fruits, Vine, Vegetables, Livestock products, Fibres and Others. For index numbers relating to India, the F.A.O. omit such commodities as are not grown in the country. They make allowance for the utilization of crops for animal feed to avoid double counting.

The weighting system, which is under scrutiny, is based on wheat relative price weights. The commodity weights are world prices. The price of a metric ton of wheat in the 1934-1938 period was 100 gold francs and, therefore, calculating each commodity price in terms of gold francs per metric ton automatically converts the prices to wheat relative weights.

General Remarks.

In constructing the indices of agricultural production, the F.A.O. is interested in inter-country comparisons. Hence it uses world price relationships in its weighting system. The objective of the other three national series is to study variations in agricultural production—for each commodity, group and for all-commodities—over a period of time. For this purpose, the index numbers should be calculated as close to farm level as possible. As such, it would appear advantageous to use current-year weighting. These index numbers at present

relate to gross production. The question of obtaining a series of net agricultural production corresponding to that of 'value added' in the national income analysis also needs to be given serious consideration.

One more point deserves to be considered. It was said in the last Chapter that the coverage of statistics of agricultural production has been increasing in recent years and that the traditional method of estimation of yield is progressively giving place to the random sampling method. Because of these factors, the production data for recent years are not strictly comparable with those for the past; a major break in comparability, owing to the above-mentioned factors, occured in 1949-50. Unless, therefore, the data for the past years are 'corrected' so as to be made comparable with those for the comparatively more recent years, index numbers based on incomparable data would be misleading. This explains, to a large extent, why the Ministry of Food and Agriculture series has been suspended; this also cautions one against using the other series for a time study. If correction of past years' data is not considered feasible on scientific grounds, the other alternative would be to start fresh compilation of the series with 1949-50 or with the average of such other years as the base after which the extension in coverage has either been negligible or can be allowed in past years' data; and, as for the other factor affecting comparability, *viz.*, changes in the method of estimation, correction factors will have to be used to make the data for the years prior to such changes, in each State and for each crop, comparable with those after the changes.

It appears rather unnecessary to have this index compiled by different agencies under, or connected with, the Government. It would seem proper if its compilation is done only in the Ministry of Food and Agriculture which is responsible for the issue of all agricultural production estimates.

The table below gives the available general index numbers of agricultural production in India as compiled by different agencies.

General Index Numbers of Agricultural Production.

		1947-48	1948-49	1949-50	1950-51	1951-52
R.B.I. (1)	....	103	107	115	111	
E. Economist (1)	....	100	95	100	94	94
F.A.O. (2)	....	99	98	102	100	

(1) Base—1936-37 to 1938-39.

(2) Base—1934-1938.

II. INDUSTRIAL PRODUCTION

Two series of index numbers of industrial production in India are available. One is the official series called Interim Index of Industrial Production compiled by the Directorate of Industrial Statistics under the Ministry of Commerce and Industry and the other is a non-official series compiled by the *Eastern Economist.*

Interim Index of Industrial Production (Ministry of C. & I. Series).

This index is issued every month for each industry and for all industries combined. The base year of the series is the year 1946. The index is based on the *Monthly Statistics of Production of Selected Industries.* The industrial classification followed in the Census of Manufacturing Industries has been adopted for the purposes of the index, since the index has been weighted by using census statistics. Weights have been allotted to the different industries in proportion to the values added by manufacture by them; these values were obtained from the *Report of the First Census of Manufactures, 1946.* In the case of coal, the gross ex-factory value at pit head of coal mined has been taken as an estimate of value added. The value added by manufacture has been taken in preference to the gross value of output as the basis of weighting, since it gives more accurate picture of the effort put forth and the value created by the industry. The index is a simple weighted arithmetic mean, calculated by the formula:

$$I = \frac{\Sigma P_1 W_1}{\Sigma W_1}$$

where P_1 is the production relative for the month, and W_1 the weight allotted to the industry.

Index numbers for all industries except sugar are adjusted for variation in the number of days in the month. Sugar industry is subject to prominent seasonal variations; adjustment is, therefore, made in the index for the seasonal variation in its output. On the basis of the value added by the sugar industry in the base year, it is entitled to an average weight of 3.54 for the 12 months. During the 8 months, April to November, the weights are less than 3.54; accordingly the deficit in the weight below 3.54 is distributed *pro-rata* among other industries. During the 4 months, December to March, when the weight exceeds 3.54, *pro-rata* reduction is made in the weights for the other industries.

The table below gives the distribution of weights for each industry or group of industries:—

Weights in Official Index of Industrial Production.

Industry or Group	Weight assigned
1. Coal	11.95
2. Sugar	3.54
3. Paints & Varnishes	0.61
4. Cement	0.67
5. Glass	0.55
6. Refractories	0.48
7. Plywood	0.15
8. Paper and paper board	1.46
9. Matches	1.21
10. Cotton textiles	43.49
11. Woollen manufactures	1.38
12. Jute	16.53
13. Chemicals	3.10
14. Non-ferrous metals	2.14
15. Steel	7.16
16. Bicycles	0.11
17. Sewing machines	0.02
18. Electric lamps	0.04
19. Electric fans	0.35
20. General Engineering & Electrical Engineering	5.06
Total ..	100.00

The largest weight is thus given to cotton textiles (cloth and yarn) with jute manufactures as a distant second.

The following table gives an idea of the progress made by industrial production in India during the last few years:

Interim Index of Industrial Production.

(Base: 1946=100)

Year		General Index
1947	..	97.4
1948	..	108.9
1949	..	106.3
1950	..	105.2
1951	..	117.4
1952	..	128.5
Apr. 1953	..	141.2

It will be seen that over-all industrial production declined in 1949 and 1950 and that thereafter it has made steady progress. In fact, in the immediate post-war years, the industries were rehabilitating themselves and were replacing old, worn-out and obsolete machinery by the new. They were in transition. That explains why the series of index numbers is called 'interim' series. The industries seem to have set up a new norm now; the results of several censuses of manufactures are also available. It should be possible now to choose one of the recent years as the base and to finalise the scheme for the construction of the index. That the index, at present, does not measure the entire industrial activity in the country would be clear from the nature of industrial production statistics available in India, discussed in the previous chapter.

The Eastern Economist Index.

The *Eastern Economist* also compiles and publishes an index of Indian industrial production. It is a weighted index. It is a monthly index; the annual index is compiled for the financial year. The industries covered, the distribution of

weights and the index numbers during the last few years are as under:—

Eastern Economist Index of Indian Industrial Production.

Year ended August 1939=100

Industry	Weights	1948-49	1949-50	1950-51	1951-52	Dec. 1952
TEXTILES						
Indian Cotton Consumption	40	119.2	98.3	90.0	99.8	121.0
Jute Manufactures	17	99.1	89.5	80.0	88.1	89.4
Total Textiles	57	112.5	95.5	86.0	96.2	110.6
FUEL AND ENERGY	10	143.2	146.8	160.8	152.7	194.7
OTHERS						
Steel ingots	8	122.2	138.6	144.7	152.6	168.9
Pig Iron	7	85.6	90.3	90.7	102.1	105.0
Paper	1	142.0	152.0	164.9	194.8	210.7
Matches	2	82.1	88.4	91.3	96.1	109.4
Paints	1	101.4	104.0	108.5	121.4	126.0
Sulphuric acid	1	92.0	111.5	116.1	118.3	120.2
Cement	3	102.2	107.5	160.8	195.4	227.9
Sugar	10	152.9	150.0	145.0	155.0	165.0
Total Others	33	115.0	125.0	127.3	138.5	149.8
GENERAL INDEX	100	115.9	109.5	104.0	115.4	129.3

According to the *Eastern Economist* index, Indian industrial production in organised industries was at an all-time peak in December, 1952. Similar position was revealed by the official interim index also.

III. INDUSTRIAL ACTIVITY

The *Capital*, an economic weekly of Calcutta, compiles and issues an Index Number of Indian Industrial Activity.

"Capital" Index of Indian Industrial Activity.

This index is compiled on monthly basis with the year 1935 as the base. It is being issued by the *Capital* since March, 1938.

The series originally selected and the weights assigned to each item for computing this index are as below:—

Weights in the Capital Index of Indian Industrial Activity.

Series Selected	Weight
Industrial Production—	
1. Indian Cotton consumption	9
2. Jute Manufactures	6
3. Steel Ingots	5
4. Pig Iron	8
5. Cement	5
6. Paper	3
Mineral Production—Coal	7
Rail and River-borne Trade	24
Financial Statistics—Cheque Clearances ..	20
Trade, Foreign and Coastal—	
Exports	4
Imports	3
Shipping, Foreign and Coastal—	
Tonnage entered	3
Tonnage cleared	3

Since March, 1941, 'Trade, Foreign and Coastal', and 'Shipping, Foreign and Coastal', have been left out. Instead, 'Notes in Circulation' (base: April, 1935 to March, 1936) with weight 6 and 'Consumption of Electricity' with weight 7 have been included. 'Rail and River-borne Trade' was first replaced by 'Railway Earnings'; since April, 1952, 'Railway Earnings' has been substituted by 'Total Number of Wagons Loaded', the weight remaining unchanged. The index for Cement remained suspended between 1938-39 and 1946-47; since February, 1948 it has appeared again with January, 1948 as the base.

An index number (relative) is first constructed for each item. The weighted geometric mean of these relatives forms the general monthly index, and seasonal fluctuations are

eliminated by means of a twelve months' moving average. A specimen of the construction of the general monthly index is given in table 17, Chapter IX, page 131. Averages of monthly indices for 12 months, April to March, give the annual indices on financial year basis. The index numbers for each item and the general index are published in the *Capital* on annual basis since 1938-39 and on monthly basis for a few recent years. The table below gives the general monthly index:

'Capital' Index of Indian Industrial Activity.

Year/Month	(Base: 1935=100) General Index
1947-48	107.0
1948-49	107.6
1949-50	109.4
1950-51	110.8
1951-52	127.4
Jan., '53	135.5

The over-all industrial activity as measured by the *'Capital'* index has been showing a consistent improvement during the last several years.

IV. COMMODITY PRICES

Index numbers of prices constitute by far the most important economic barometers. Prices themselves are of several categories. Harvest prices, wholesale prices and retail prices are the more important of them; and, it is for these categories of prices that index numbers are available in India. Since retail prices in urban areas cannot be said to be representative of those in the rural areas, separate indices for the two areas are also now available. The following index numbers relating to prices are discussed here:

1. Index Numbers of Harvest Prices compiled by the Directorate of Economics and Statistics:
2. Index Numbers of Wholesale Prices compiled by the Office of the Economic Adviser, Ministry of Commerce and Industry;

3.* Index Numbers of Wholesale Prices in Calcutta compiled by the Department of Commercial Intelligence and Statistics;
4. Index Numbers of Retail Prices (Urban Centres) compiled by the Labour Bureau;
5. Index Numbers of Retail Prices (Rural Centres) compiled by the Labour Bureau.

Index Numbers of Harvest Prices.

Index Numbers of Harvest Prices of Principal Crops are compiled in the Directorate of Economics and Statistics, Ministry of Food and Agriculture. For the period 1939-40 to 1949-50, they were published in the *Indian Agricultural Price Statistics, 1950-51.* The prices data forming the basis of these indices cover Part A States of India and Delhi and are obtained from non-official agencies, principally the branches of the Imperial Bank of India. The prices are the averages of the weekly quotations during the harvesting period at the important marketing centres adjoining the major producing areas for each crop.

The commodities covered by the index are:—Rice, wheat, jowar, bajra, barley, maize, gram, linseed, sesamum, rape and mustard, groundnut, sugar (raw), cotton, jute and tobacco. The harvest prices ruling in 1938-39 constitute the base year prices. Chain base method of construction of index numbers is used, in view of the fact that the price quotations are not uniformly available throughout the period. Current year's price relatives are computed in relation to the previous year and linking is done with the base year through the intermediary years. The average price relative is worked out for each State in respect of each commodity as follows:—

(*i*) For each year, the price relative of each variety of a commodity at each centre is computed in relation to the corresponding price for the previous year (equivalent to 100).

(*ii*) A simple geometric average of such price relatives for the different varieties quoted at each centre

* The Bombay Wholesale Price Index Number, which was compiled by the Bombay Labour Office with July, 1949 as the base and with 40 items, was discontinued in November, 1944.

gives the price relative of the commodity at that centre.

(*iii*) A simple geometric average of the different price relatives at the different centres gives the single price relative for the State for the commodity.

The price relatives for the different States are combined into a commodity index by taking their weighted geometric mean, the figures of current year's production in the different States being used as weights. As the price relative in each year is computed relative to the previous year, the system of moving weights has been adopted.

After compiling the price indices in each year relative to the previous year, these have been linked through successive intermediate years with the base year price relatives, to get the index numbers for each of the commodities.

The weighted geometric mean of the index numbers for these 15 commodities, with the average value of their production during the three years ending 1938-39 as weights, gives the All-India Index Number of Harvest Prices. The table below gives the commodity indices and the all-commodities indices for 1939-40 and the latest two years:—

Index Numbers of Harvest Prices.

(Base: 1938-39=100)

Group or Commodity	1939-40	1948-49	1949-50
Foodgrains Gr. ..	117	500	546
Oilseeds Gr. ..	121	538	635
Sugar, raw ..	82	213	365
Cotton ..	143	322	365
Jute ..	124	718	622
Tobacco ..	123	506	425
All Commodities ..	112	453	514

It will be seen that by 1948-49, harvest prices had risen 4½ times over those in 1938-39 and that they were more than 5 times as high in 1949-50.

Economic Adviser's Wholesale Prices Index.

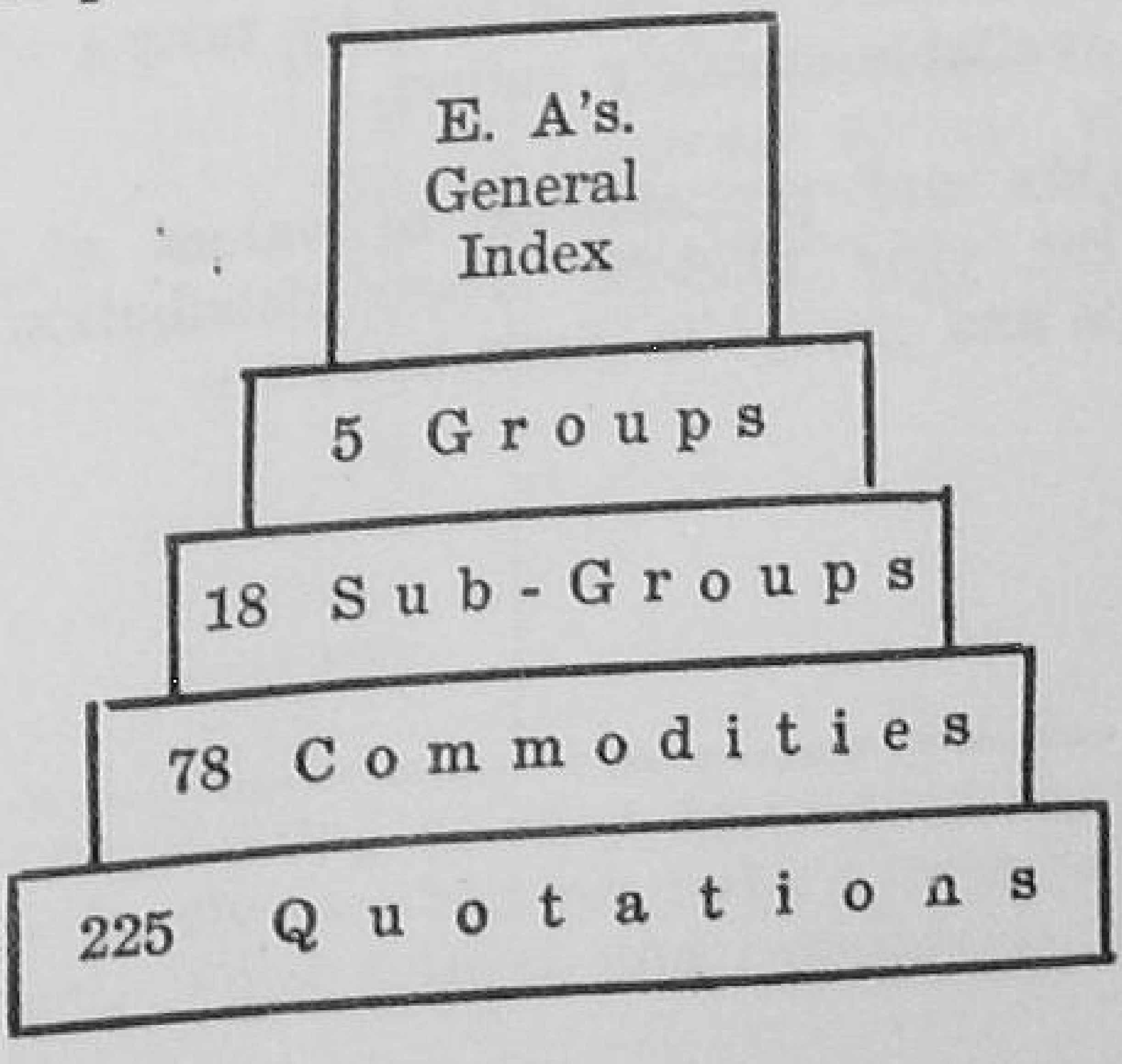
For a study of the price position in India in recent years recourse has to be taken to the Economic Adviser's Index Numbers of Wholesale Prices with the year ended August 1939 as the base. These index numbers are available for five groups *viz.*, (I) Food Articles, (II) Industrial Raw-Materials, (III) Semi-manufactures, (IV) Manufactures and (V) Miscellaneous. Each of these group index numbers is composed of certain sub-groups and each sub-group comprises a few commodities of allied nature, as will be seen from the table below:

Group	*Sub-Group*	*Commodity*
(I) Food Articles.	(a) Cereals	(1) Rice (2) wheat (3) jowar (4) bajra.
	(b) Pulses	(1) Gram (2) dal arhar.
	(c) Others	(1) Tea (2) coffee (3) sugar (4) gur (5) salt.
(II) Industrial Raw-materials.	(a) Fibres	(1) Cotton raw (2) jute raw (3) silk raw (4) wool raw.
	(b) Oilseeds	(1) Groundnut (2) linseed (3) gingelly seed (5) rape seed (6) cotton seed (7) copra.
	(c) Minerals	(1) Coal (2) manganese ore (3) mica (4) iron ore.
	(d) Others	(1) Hides raw (2) skins raw (3) lac (4) rubber.
(III) Semi-manufactures.	(a) Leather	(1) Cow hides (2) buffalo hides (3) sheep skins (4) goat skins.
	(b) Mineral oils	(1) Kerosene oil (2) petrol.
	(c) Vegetable oils	(1) Coconut oil (2) groundnut oil (3) mustard oil (4) castor oil (5) gingelly oil (6) linseed oil.
	(d) Cotton yarn (grey) Indian.	—
	(e) Metals	(1) Pig iron (2) semis (3) zinc spelter (4) brass (5) copper.

	(f) Oilcakes	(1) Groundnut cake (2) coconut cake (3) gingelly cake.
	(g) Others	(1) Timber (2) coir yarn.
(IV) Manufactures.	(a) Textile products.	(1) Jute manufactures (2) cotton mfs. (3) rayon and silk mfs. (4) woolen mfs.
	(b) Metal products.	(1) Iron and steel mfs. (2) machinery (3) vehicles.
	(c) Other finished products.	(1) shoe (2) tyres and tubes (3) cement (4) paper (5) matches (6) glass (7) soap (8) chemicals (9) paints (10) cigarettes (11) liquors (12) dye stuffs.
(V) Miscellaneous.	-	(1) Vanaspati dalda (2) spices and condiments (3) betel nuts (4) cashew nuts (5) bricks and tiles (6) tobacco leaf.

Thus, the five groups in all take in 78 commodities through 18 sub-groups. For each of the 78 commodities several quotations are taken, so that the total number of quotations entering the index is about 225. The five group indices are combined into an all-commodities index, commonly known as the General Index of Wholesale Prices.

The whole picture of this index is, thus, as under:—



The index is compiled every week, and is published on weekly, monthly and annual basis. It is compiled on the basis of one-day-a-week prices, on or about Friday, received from official and non-official sources. The index is published in the *Index Number of Wholesale Prices in India,* a weekly bulletin issued by the Economic Adviser. The bulletin gives the quotations for each individual commodity, commodity indices, sub-group and group indices and the general index for the current and previous weeks. A summary is issued earlier to the daily press.

For each weekly quotation in respect of a commodity, a price relative is first calculated; simple geometric average of the price relatives of the several quotations used for a commodity constitutes the **commodity index.** If only one quotation is used for a commodity, the price relative would itself constitute the commodity index. Of the commodity indices falling within a sub-group, weighted geometric average is computed to arrive at the **sub-group index.** Weighted geometric average of the sub-group indices gives the **group index.** The group index numbers are lastly combined, through the process of weighted geometric mean, to yield a single **all commodities index,** or the **general index.**

The monthly indices are obtained by taking the geometric average of the weekly index numbers. Similarly, the annual average index numbers are calculated by taking the geometric mean of the available monthly indices.

The weights are proportional to values of commodities marketed in the year 1938-39. Their distribution as between the sub-groups and groups is given below:—

Group and Sub-Group Weights in
E. A's. Wholesale Prices Index.

Group & Sub-group		Weight
FOOD ARTICLES		(31)
Cereals		59
Pulses		8
Others		33
		100
INDUSTRIAL RAW MATERIALS		(18)
Fibres		53
Oilseeds		30
Minerals		10
Others		7
		100
SEMI-MANUFACTURES		(17)
Leather		8
Mineral oils		13
Vegetable oils		16
Cotton yarn		35
Metals		18
Oilcakes		5
Others		5
		100
MANUFACTURES		(30)
Textiles		64
Jute	11	
Cotton	45	
Rayon & silk	6	
Woollen	2	
Metal products		17
Others		19
		100
MISCELLANEOUS		(4)
All Commodities		(100)

Note.—Figures in brackets indicate Group Weights, and figures without brackets indicate Sub-group Weights within the respective Groups.

These weights, relating as they do to 1938-39, are old and out of date now in the context of India's post-war and post-partition economy. Food articles and raw materials get about

half the total weights in the index; manufacturing industries which, in fact, grew during and after World War II get relatively less weightage. The manner of selection of weights has also been criticised often: the weights are based on the gross market value of commodities and not on the net addition to their value in the process of production. For instance, raw-cotton and raw jute are first entered at their marketable value and, later, when cotton cloth and yarn and jute manufactures are entered, they are counted at their full values. This procedure involves an element of duplication in weights, resulting in greater importance to commodities like cotton and jute textiles as compared with commodities like cement or coal. This deficiency can be made up now that the results of the census of manufacturing industries are available for several years. Again, while the export commodities are weighted in proportion to the total quantity produced (quantities exported being ignored), the imports into the country are not represented in the index. Thus the commodities included in the index do not give a fair sample of the supplies available, or exchanged, for consumption in the country. For these limitations of it, the E. A's index does not necessarily show changes in the prices of different commodities in proportion to the importance of each commodity to the present national economy.

The E. A's index has also been criticized on the ground that the number and character of quotations used in its construction are not representative. For instance, whereas the number of quotations used for rice is three that for shoes, a comparatively less important commodity, it is eight. Wheat with a weight of 3.7% has only 3 quotations, while tyres and tubes with a weight of only 0.3% are represented by as many as 9 quotations. As regards the character of quotations, open market prices of sugar were not taken into account even when 'dual' price system was in force, or when control over its price, sale, etc. was removed.

The 'food articles group index' is very often referred to as 'food index', and the term 'food' is taken by many as a synonym for cereals or foodgrains, whereas the 'food articles group' really includes not only cereals but also gram, *arhar* pulse, tea, coffee, sugar, *gur* and salt. Cases have not been found wanting when, if the 'food articles group index' showed a rise, it was

made out in uninformed quarters that cereals prices had gone up. There is, therefore, a clear need for placing before the public a 'cereals' index as distinct from the 'food articles group index'.

For the cereals sub-group in the index only 8 quotations are used: 3 for rice, 3 for wheat, and 1 each for jowar and bajra. This number is obviously too small not only because of the vastness of the country but also because of relaxations in food controls which started in the middle of the year 1952 and the consequent restoration of intra-state trade.

In view of what has been said above, there is a clear case for overhauling this index. The number of quotations needs to be increased, the weights to be re-assessed, the base year to be revised forwards, the grouping to be altered so as to give each distinct group a proper place, more commodities to be brought in, and so on. The aim should be to make it really a general purpose index, a measure of general price level.

The table gives index numbers for the five economic groups and the all-commodities index, since 1947:—

E. A's Index Numbers of Wholesale Prices.

(Base: Year ended August 1939=100)

Year	Food Articles	Industrial Raw Materials	Semi-Manufactures	Manufactured Articles	Miscellaneous Articles	All Commodities
1947	292.2	365.3	251.9	276.6	468.1	297.4
1948	374.1	430.5	316.6	340.6	504.2	367.1
1949	389.2	463.9	328.1	343.9	545.6	381.1
1950	410.4	503.2	340.9	348.4	683.1	400.7
1951	409.8	608.2	377.5	395.5	729.8	439.3
1952	359.8	453.8	346.5	377.5	635.6	386.9
Week ended 12 Sep. '53.	400.5	484.3	366.6	363.3	718.5	405.9

The effect of devaluation which took place in the latter half of 1949 and that of the Korean War which broke out in the middle of 1950, on the general price level in the country is clearly visible from the above table. The policy of credit

restriction adopted by the Reserve Bank of India towards the end of 1951 brought about an all-round decline in prices in 1952, the year's index being about 50 points lower than the average for 1951.

E. A's General Purpose Vs. Sensitive Index.

The current index number of wholesale prices, discussed above, is intended to be a general purpose index. This series is available from January, 1947. During the War, the Economic Adviser had initiated a sensitive wholesale price index number also, with the week ended 19th August, 1939, as the base. This sensitive series was worked out as a simple geometric average of the price changes of 23 selected commodities, for which 46 quotations were utilised. This series is available from 1939 to December, 1947, when it was discontinued.

Thus the sensitive and general purpose series are available for a common period of 12 months, January, 1947 to December, 1947, which enables splicing of the all-commodities index of one series with the other.

Index Numbers of Wholesale Prices in Calcutta.

The Index Numbers of Wholesale Prices in Calcutta are compiled in the Department of Commercial Intelligence and Statistics, Ministry of Commerce and Industry, and are regularly published in the *Indian Trade Journal*.

This index number is a monthly index and is compiled from 72 items which are divided in the following 16 groups:—

Group	Number of Items
1. Cereals	8
2. Pulses	6
3. Sugar	5*
4. Tea	3
5. Other food Articles	9
6. Oilseeds	3
7. Oil, mustard	2
8. Jute, raw	3
9. Jute manufactures	4
10. Cotton, raw	2

11.	Cotton manufactures	7*
12.	Other textiles (Wool and silk) ..	2
13.	Hides and Skins	3
14.	Metals	6
15.	Other Raw and manufactured Articles	8
16.	Building Materials (Teak wood) ..	1
	Total ..	72*

A separate index number is worked out for each group. The index number of each group is calculated by computing the simple arithmetic average of the price relatives of the articles included in the group. Though the index number of each individual group is calculated by the use of simple arithmetic average, an element of weighting is introduced by taking more than one quotation for some items within a group *e.g.* "cereals" includes 4 varieties of rice and only one variety each of wheat, barley, maize and oats. The base for the index is July, 1914.

The general index number is arrived at by calculating the simple arithmetic average of all individual price relatives included in the compilation. This general index number is weighted implicitly, the weights in each case being equal to the number of items included in each group.

Owing to the market quotations for a number of articles not being available for some time past, the Department of Commercial Intelligence and Statistics reviewed the question of continuing the series and decided that the compilation of the index should be continued on the present basis as a purely temporary measure. It is felt that under the existing economic conditions, this index cannot be treated as necessarily having an all-India significance, as it relates to Calcutta prices only. And, as it is based on one-day-a-month prices, it is not sufficiently representative of the average monthly price level. Distribution of weights in this index is also not representative of all-India. For instance, rice is given a weightage of 4 and wheat that of 1. This may be true of Calcutta but not of all-India. The index, therefore, has limited utility.

*Two items have been dropped under sugar since February, 1934 and one item under cotton manufactures since September, 1936 as their quotations are no longer available and suitable substitutes cannot be found for them. Thus the total number of items included in the index has come down to 69.

The table below gives the all-commodities index for 1939 and the latest 3 years:

Index Numbers of Wholesale Prices in Calcutta.
(Base: July, 1914=100)

Year	Index
1939	108
1950	511
1951	541
1952	410
Jan. 1953	399

Index Numbers of Retail Prices (Urban Centres).

The Index Numbers of Retail Prices in 18 selected Urban Centres in India are compiled in the Labour Bureau, Ministry of Labour, and are regularly published in the *Indian Labour Gazette*. These are unweighted index numbers. They are compiled on monthly basis for three important groups: All Articles of Food (with three sub-groups, *viz.* cereals, pulses and other articles of food), Fuel and Lighting, and Miscellaneous. The data for these indices are collected weekly. Originally the base for some of these indices was January 1943, and for some others different. Now, they are being issued on a uniform base: 1944=100. The 18 selected centres are situated in 5 States, *viz.*, Bombay 3, Punjab 1, U.P. 5, Bihar 1, and West Bengal 8. The table below gives the group indices for one centre each in the 5 States:

Unweighted Retail Price Index Numbers by Groups.

(Base: 1944=100)

Centre	All Articles of Food		Fuel and clothing		Miscellaneous	
	January		January		January	
	1952	1953	1952	1953	1952	1953
Bombay						
Surat	150	132	98	103	163	157
Punjab						
Amritsar	165	145	108	113	173	164
U.P.						
Lucknow	171	155	108	96	201	158
Bihar						
Patna	197	164	130	131	169	157
West Bengal						
Calcutta	168	139	104	100	167	156

Index Numbers of Retail Prices (Rural Centres).

The Index Numbers of Retail Prices in 12 selected rural centres in different parts of India are compiled in the Labour Bureau, Ministry of Labour and are regularly published in the *Indian Labour Gazette.*

These "unweighted" index numbers of retail prices for each of the centres are compiled on monthly basis for four important groups: 1. All Food Articles (including the sub-groups, cereals, pulses and others for which separate indices are worked out); 2. Fuel and Lighting; 3. Clothing; and, 4. Miscellaneous. All the centres are railway stations. The prices of articles entering the group are collected weekly in each centre by the Station Masters and their work is regularly supervised by the Inspectors of Railway Labour within whose beat the stations lie. The returns are sent to the Labour Bureau, who compile the monthly index numbers.

Originally, the compilation of the retail price index numbers was started with the prices ruling in January, 1943 as base but later, in order to secure comparability of indices as between the rural and urban centres, these were brought to a common

base, *viz.*, January to December, 1944=100. This change in the base period was introduced from January, 1946.

The following table gives the group indices for each centre for the month of January 1952 and 1953:

Unweighted Retail Price Index Numbers by Groups.

(Base: 1944=100)

	Food Articles		Fuel and Lighting		Clothing		Miscellaneous	
	January		January		January		January	
	1952	1953	1952	1953	1952	1953	1952	1953
Eastern Zone								
1. Bamra	162	134	102	107	187	216	193	193
2. Maibang	156	....	141	....	163	....	125	....
3. Shankargarh	184	160	101	97	164	141	141	144
4. Sonaili	237	183	125	125	264	199	332	278
Northern Zone								
5. Multapi	164	161	108	133	167	163	148	196
6. Nana	252	212	144	117	244	205	208	189
7. Salamatpur	271	236	124	188	132	136	258	268
Southern Zone								
8. Krishna	213	166	141	115	171	138	213	176
9. Lakh	212	182	86	86	129	138	144	131
10. Malur	167	....	89	....	406	....	176	....
11. Muniguda	308	218	105	128	207	194	241	199
12. Kudchi	176	164	104	127	155	122	202	191

These indices roughly measure the fluctuations in the prices of certain groups of consumption items as compared to the base.

V. SECURITY PRICES

Index Numbers of Security Prices in India were first compiled by the Economic Adviser to the Government of India, Ministry of Commerce and Industry. They were discontinued with effect from the close of December, 1949. The Reserve Bank of India started a weekly series of these index numbers in January, 1946, with the year 1938 as the base. In April, 1953, the Reserve Bank discontinued this meekly series and started a revised one, with 1949-50 as the base.

Reserve Bank Index of Security Prices (Old Series).

The index numbers of industrial securities issued by the Reserve Bank were intended to measure the price changes in important industrial sectors from the point of view of general economic situation.

The method of computation followed was to select a large and representative sample of important scrips in each industry and calculate every week, an index for that industry (sub-group) by taking an unweighted geometric average of the price relatives. The indices for different sub-groups were then combined by taking a weighted arithmetic average to obtain a group index, the weights being proportional to the total paid-up capital of all companies whose shares were quoted in the concerned Stock Exchanges in 1946. The chain base method was used; each group and sub-group index was a chained index obtained by computing link relatives each week and chaining to the previous week's index. Thus, ultimately the indices were based on the calendar year 1938, and, with that base, published regularly in the *Reserve Bank of India Bulletin.*

The pre-War base had become out of date; some of the new loans floated by the Government of India and State Governments had begun to be actively floated in the market and, therefore, required to be included in the index; a number of big concerns had come to be set up and called for their inclusion in the index; the weights relating, as they did, to 1946 could not claim to be representative of the changed conditions after Independence; the method of computation also required some improvements. These circumstances led to the emergence of the Revised Series in August, 1953.

Reserve Bank Index of Security Prices (Revised Series).

Though the revised series was for the first time published in August, 1953 issue of the *Reserve Bank of India Bulletin,* it was made available with effect from the first week of April, 1953. For the period April to July, 1953 the old series is also available. A comparison of the two series for this common period of four months shows that their trends are more or less parallel.

The principal features of the revision may be grouped under six heads:—

2. A number of new scrips has been added and a few old year 1938 to the financial year 1949-50.

2. A number of new scrips has been added and a few old ones dropped; for Variable Dividend Industrial Securities, quotations from the Delhi Stock Exchange have been included; new series of index numbers for Debentures have also been constructed.

The scrips for the old series were mainly selected from those which were traded during the base year 1938; originally, about 400 scrips were selected of which a few had to be dropped when they ceased to be traded. The scrips for the revised series have been selected on two criteria: (1) the size of the concern as judged by the paid-up capital and (2) the activity of the scrip in the Stock Exchange. Thus, most of the large concerns are included irrespective of the activity of their shares on the Stock Exchange. The amounts outstanding and the dates of maturity of the scrips have, in addition to the activity of the scrip on the Stock Exchange, been taken into account in selecting Government and Semi-Government Securities and Debentures. Very short-dated loans were excluded; as, however, the revised series is a general purpose one, a few short-dated loans of maturity period ranging from 2 to 5 years have been selected. Debentures too have been selected on similar basis. State Loans relating to Part 'B' States were already there in the old series; a number of those for Part 'A' has been added in the revised series. All the Government of India Securities selected for the revised series are common to the three Stock Exchange Centres, Calcutta, Bombay and Madras. A few variable dividend scrips traded on the Delhi Stock Exchange have been included in the revised series in order to give proper representation to the companies situated in northern India. For this centre, scrips belonging to the two groups, Government and Semi-Government Securities and Debentures, could not be selected, as they are not traded there. So were not included the Preference Shares too. The point to remember, therefore, is that while the all-India indices for Variable Dividend Securities relate to 4 centres in the revised series, the indices for the three other main groups relate to only 3 centres. Another significant feature of the revised series is that a new main group for Debentures has been added

in order to secure a complete coverage of all the various types of securities transacted on the Stock Exchanges.

3. Some minor alterations have been made in the grouping of the scrips.

Now, there are 4 main groups against 3 formerly, and 25 groups, as shown below:—

Main Groups and Groups in Reserve Bank Index of Security Prices.

Main Group	Group
A. Government & Semi-Government Securities.	1. Government of India Loans, 2. Indian State Loans, 3. Municipal, Port Trust & Improvement Trust Loans.
B. Debentures of Industrial Concerns.	1. Electricity, 2. Jute, 3. Plantations, 4. Sugar, 5. Coal, 6. General Engineering, 7. Shipping, 8. Others.
C. Preference Shares.	1. Electricity, 2. Cotton, 3. Jute, 4. Coal, 5. Iron & Steel, 6. Sugar, 7. Paper, 8. Plantations, 9. Miscellaneous.
D. Variable Dividend Securities.	1. Financial concerns, 2. Public Utilities, 3. Manufacturing industries, 4. Non-Manufacturing industries, 5. Miscellaneous.

Main group C, Preference Shares, is the same as the Fixed Dividend Industrial Securities main group of the old series. The five groups under the main group D are divided into 18 sub-groups.

4. Adjustments have been made for the sudden decline in the prices of scrips owing to the issue of bonus shares or issue of new capital on bonus terms.

5. In respect of all industrial securities, the weights now used are proportional to the market values of shares during the base period instead of those proportional to their paid-up capital. Market value correctly represents the markets' appraisal of relative importance of different industries. This system of weighting is followed in many foreign countries, such as the U.S.A., France, New Zealand, Denmark, Norway and Finland.

As in the old series, the weights used for combining the revised sub-group indices of Government and Semi-Government Securities are proportional to the average outstandings of all the securities traded at the respective centres during the base year. The sub-group index numbers continue to be unweighted ones.

6. The chained index numbers of the sub-groups are weighted, instead of weighting the link-relatives.

The weekly price of each scrip is obtained by averaging daily quotations. Price relatives for the different scrips are then calculated by dividing the weekly average price of each scrip by the corresponding average price for the immediately preceding week. The unweighted geometric average of the price relatives of the scrips included in the sub-group at each centre is the link relative for the sub-group at that centre. These link relatives are then forged into a chain by serial multiplication to get the sub-group index for that centre. The regional main group indices are arrived at by computing the weighted arithmetic averages of the sub-group indices of the region. The all-India sub-group indices are the weighted arithmetic averages of the regional sub-group indices and the all-India main group indices are the weighted arithmetic averages of the all-India sub-group indices. This method is somewhat different from the one followed in the old series: formerly the different sub-groups were combined by taking a weighted

arithmetic average of sub-group link relatives and then chaining the link-relatives thus obtained. The revised method avoids the minor bias that might have crept into the main group indices, were the dispersions of the sub-group indices large.

The table below gives the month-end main-group indices of the revised series:

All-India Index Numbers of Security Prices.

(Base: 1949-50=100)

Main Group	April 25 1953	May 30 1953	June 27 1953	July 25 1953
A. Govt. & Semi-Govt. Securities.	90.5	90.7	90.5	90.4
B. Debentures of Industrial Concerns.	100.7	100.5	100.6	100.5
C. Preference Shares	86.3	87.0	87.3	87.4
D. Variable Dividend Industrial Securities.	92.8	94.9	93.9	93.2

As compared with the base year, the main group A, C, and D, showed a fall in April-July, 1953, while main group B recorded a nominal rise. The fall was the greatest in the case of Preference shares.

VI. WORKING CLASS COST OF LIVING INDEX NUMBERS

The following series of working class cost of living index numbers in India are regularly published in the *Indian Labour Gazette:*

(1) Labour Bureau Working Class Cost of Living Index Numbers for 15 centres;

(2) Labour Bureau Interim Series of Cost of Living Index Numbers for Mercara (Coorg) for Plantation Workers;

(3) Working Class Cost of Living Index Numbers for 13 centres compiled by the State Governments concerned;

(4) Labour Bureau Interim Series of All-India Average Working Class Cost of Living Index Number.

Labour Bureau Index Numbers for 15 Centres.*

The base year of these index numbers is 1944. They are compiled on monthly basis in the Labour Bureau, Ministry of Labour, for the five usual groups, besides the general index. For each of the 15 centres, the various items of expenditure included in the index and the weights have been determined on the basis of family budget enquiries conducted during the period 1943-1945. The table below gives a comparative picture of the weights for each group in respect of some of the 15 centres:

Weights in Labour Bureau C/L Indices.

Centre	Food	Fuel and Lighting	House Rent	Clothing Bedding & Footwear	Miscella-neous	All Items
Delhi	61.0	8.8	6.4	10.8	13.0	100.0
Jamshedpur	65.8	5.4	4.7	10.2	13.9	100.0
Cuttack	67.1	8.9	3.0	8.1	12.9	100.0
Gauhati	63.4	8.6	4.5	5.1	18.4	100.0
Ludhiana	61.2	9.2	3.7	10.8	15.1	100.0
Jabalpur	58.2	7.2	2.7	12.6	19.3	100.0
Kharagpur	69.2	5.9	3.4	10.6	10.9	100.0

It will be seen that food items account for roughly 60 to 70 per cent. of the total working class expenditure, and that the miscellaneous group calls for greater expenditure than any of the three groups other than food. It is clear, therefore, that a given percentage rise in the prices of food items will have a greater impact on the general index than the same order of increase in the prices of other groups. This is also true of the miscellaneous group *vis-a-vis* the non-food groups. Food items, however, do not include cereals alone, but many other items of

* A 16th centre, viz., Bhopal city, has since been added with effect from January, 1952 and with base: January to December 1951=100.

direct consumption like salt, sugar, *ghee*, spices, vegetables etc. The items included in each group in the different centres being dependent on the family budget enquiries, they are not naturally common as between different centres. They can no doubt be classified into similar groups uniformly, and the same has been done.

The table below gives these index numbers for 1945 and the latest two years, for each of the 15 centres:—

Labour Bureau Working Class Cost of Living Index Numbers.

(Base: 1944=100)

Centre	1945	1951	1952
1. Delhi	103	142	143
2. Ajmer	110	178	174
3. Jamshedpur	100	160	154
4. Jharia	97	184	175
5. Dehri-On-Sone ..	99	197	188
6. Monghyr	105	188	172
7. Cuttack	102	181	160
8. Berhampur	101	190	156
9. Gauhati	90	141	142
10. Silchar	92	159	141
11. Tinsukhia	94	124	128
12. Ludhiana	105	167	151
13. Akcla	98	165	156
14. Jabalpur	95	168	150
15. Kharagpur	97	136	133

The above table gives a measure of the over-all changes from the level in the year 1944 of retail prices of goods and services that entered into working class expenditure during the period of family budget enquiries. It shows, for instance, that an average working class family in Delhi was required to spend 43% more than it did during 1944 to maintain the base-year standard of living. The table also shows that while the cost of living in 1951 and 1952 was definitely higher than that in 1945 in all the 15 centres, it was, in most cases, lower in 1952 than in 1951.

Labour Bureau Index Number for Plantation Workers.

This index number is being compiled, as an interim measure, for plantation workers in Mercara (Coorg) only. The base period for this index is July to December, 1948. The family budget enquiry at this centre was conducted during July to September, 1947. The grouping of the items of expenditure, method of construction, etc. are the same as in the case of the Labour Bureau working class index numbers for 15 centres. The general index stood at 111 in 1949, 118 in 1951 and 110 in 1952.

States' Cost of Living Index Numbers.

Working class cost of living index numbers in 13 industrial centres situated in 7 States are being regularly published in the *Indian Labour Gazette*. Besides these, the State Governments compile cost of living index numbers for several other centres and classes of population. The States publish their index numbers in their own *Labour Gazettes* or *Bulletins*.

These index numbers are not directly comparable with one another because of the diversity in their scope and construction. The base period of the 13 indices published in the *Indian Labour Gazette* varies both in time and length: the base period varies from 1926-27 in the case of Ahmedabad to 1943-44 in that of Hyderabad; the length of the base period varies from just a month in 5 centres to a year in the other 8. Likewise, the weights are also based on family budget inquiries conducted during varying periods, nearabout the base period in each case; these index numbers, therefore, refer to widely differing standards of living. The frequency with which the price quotations are collected is also not uniform, being a week in some cases and a month in others. The coverage or scope of the items included in these indices is also not uniform; certain groups, particularly the miscellaneous group, in some indices are not as comprehensive as in others. The agency compiling these index numbers is generally the Labour Department; in Madras it is the Office of the Economic Adviser in the State.

The method of construction of these indices, though not quite uniform, follows a certain general pattern. The items in the list of articles consumed by the working classes are grouped under the usual 5 heads; separate indices are worked out for

the individual groups, the index for any group being the weighted average of the price relatives of the various items in the group. The weight assigned to any item in a group represents the ratio which the expenditure on that item bears to the total expenditure on all items included in the group. The group indices are similarly combined into a general index by weighting each group in the proportion of total expenditure incurred on each group. Arithmetic average is employed in striking the weighted average.

The weightage assigned to each of the five groups in some centres is given below:

Weights in State's C/L Index Nos.

Group	Bombay	Madras	Nagpur	Kanpur
Food	47 (53)	53	64	42 (60)
Fuel & Lighting	7 (8)	8	10	6 (9)
Clothing	8 (9)	6	11	8 (12)
House-rent	13 (14)	15	15	7 (10)
Miscellaneous	14 (16)	13		6 (9)
Total	89 (100)	100	100	69 (100)

It will be seen that the weightage assigned to the Food group is the highest; this group alone accounts for more than half the total weightage and, therefore, has a greater impact on the general index than other groups for a uniform percentage variation in prices. Within the Food group, the weightage assigned to cereals is not necessarily equal to or higher than that assigned to non-cereal commodities, the ratio between the weights of cereals and non-cereals being 47:53 in Madras, 45:55 in Kanpur and 36:64 in Bombay. It follows, therefore, that even for a uniform rise in the prices of cereals and non-cereals, the impact of the latter on the Food index will be much greater than that of cereals because of the heavier weightage assigned to the non-cereals. It may even happen that the Food group index may show a rise, *despite* a fall in the prices of cereals if the rise in the prices of non-cereals is of a significant order. These facts must be kept in mind while studying the

Food group index of the cost of living index numbers. The Food group indices for the 13 centres under reference are also published regularly in the *Indian Labour Gazette.*

Most of these 13 centres have been under rationing of one type or another. For the period that foodgrains were (or, are still) supplied to working-class population from the ration shops, naturally controlled prices of foodgrains entered (or, enter) the index numbers.

Although these index numbers are compiled on different base periods they are shifted to a common base, August 1939=100, for facility of comparison. The base shifting is done by mere arithmetical conversion. The table below gives the original base-periods, and general cost of living index numbers on uniform base, in respect of the 13 centres for 1940 (wherever available), 1946, 1952 and January 1953:—

Working Class Cost of Living Index Numbers (States' Series).

(Base shifted to August, 1939=100)

Centre		Original Base	1940	1946	1952	1953 Jan.
BOMBAY—						
Bombay	....	July 1933—June 1934	107	246	321	330
Ahmedabad	....	Aug. 1926—July 1927	108	286	355	356
Sholapur	....	Feb. 1927—Jan. 1928	104	290	424	425
Jalgaon	....	Aug. 1939	109	326	447	436
U.P.—						
Kanpur	....	Aug. 1939	111	328	441	454
M. PRADESH—						
Nagpur	....	Aug. 1939	110	285	380	381
MADRAS—						
Madras	....	July 1935—June 1936	109	240	337	334
TR.-COCHIN—						
Ernakulam	....	Aug. 1939	....	305	387	386
Trichur	....	Aug. 1939	....	294	376	373

MYSORE—						
Bangalore	....	July 1935— June 1936		221	346	347
Mysore	....	—Do—	....	238	343	349
Kolar Gold Fields	....	—Do—	....	247	379	384
HYDERABAD—						
Hyderabad City	....	Aug. 1943— July 1944	....	117	164	167

The extent of variation in cost of living in the different centres can be clearly seen from the above table. By January, 1953, the greatest rise took place in Kanpur as compared with August, 1939.

We may now take up a detailed description of the scope and method of construction of the Bombay and Kanpur working class cost of living indices.

Bombay Working-Class Cost of Living Index.

This index was first published in 1921, and was based on the aggregate consumption method in the absence of any reliable weights to be given to different items. The Bombay Labour Office conducted the first inquiry into working-class family budgets in Bombay City between May 1921 and April 1922 and a second inquiry between May 1932 and June 1933 to ascertain weights proportional to the relative expenditure on the different items consumed by an average Bombay workers' family. The results of the second inquiry have been used in the compilation of the revised index, commodities have been made as comprehensive as possible and the "miscellaneous" group has been added. Average prices for July 1933 to June 1934 are equated to 100, and the index is published in the *Labour Gazette* issued by the Labour Office of the Government of Bombay.

The items included in the revised index are divided into five main groups, *viz.*, food, fuel and lighting, clothing, rent and miscellaneous. The **Food group** includes 28 articles, which are: rice, *patni*, wheat, *jowari*, *bajri*, *turdal*, gram, raw sugar (*gur*), refined sugar, tea, four varieties of fish, mutton, milk, *ghee*, salt, dry chillies, tamarind, turmeric, potatoes, onions, brinjals, pumpkins, cocoanut oil, sweet oil and readymade tea. The expenditure on other articles which can be included in the food

group has been proportionally divided among items of like nature included in the food group; for example, the expenditure on refreshments has been added to expenditure on ready-made tea, and that on sweet-meats has been divided equally between sugar and milk. **Fuel and lighting group** includes charcoal, fire-wood, kerosene oil and matches. **Clothing group** includes *dhotis*, coating, shirting, cloth for trousers, *sarees*, and *khans*. The figure adopted for **House-rent** is the average rent per tenement obtained as a result of the 1932-33 family budget inquiry. **Miscellaneous group** includes barber (shave), washing soap, medicine, *supari, bidis*, travelling to and from native place and newspapers. Thus, this index number includes 46 articles.

The price quotations for almost all the articles, except clothing articles, four varieties of fish, brinjals, and pumpkins, are collected weekly by the officers of the Labour Office from two shops in twelve different industrial areas. The prices of all the clothing articles except *khans* are obtained from four different cotton mills having retail shops in Bombay City. Prices of fish, brinjals and pumpkins are taken from the Municipal records.*

The method* adopted for computation of the index number is very similar to that of the British Ministry of Labour. The index is weighted twice in the process of construction. Price quotations for the current month are first expressed as percentages of the average prices for the base year, 1933-34. These percentages are weighted by the percentages which expenditure on each item bears to the total expenditure on the group to which it belongs, and the products are summated. In the Fuel and Lighting group, for instance, charcoal gets a weightage of 30, firewood 52, kerosene oil 16 and matches 2, totalling 100. Sum of the products divided by 100 gives the weighted average index for each group. The group index numbers are again weighted by the percentage distribution of the expenditure on

*With effect from the index for the month ending 15th May, 1943, the method of compilation of the index number for the cereals sub-group has been readjusted consequent on the introduction of rationing at Bombay in May, 1943. Prices of different cereals sold in Government grain shops and authorized ration shops in working-class localities, and shops run by employees of labour are being utilized. Besides, the indices from June, 1943, onwards relate to the calendar month and not to the month ended the 15th as before.

each of the groups, and then divided by the sum of the weights. The resulting weighted average is the general index.

The percentages by which group index numbers are weighted are those arrived at as a result of the 1932-33 inquiry. except in the case of the 'miscellaneous' group whose weight is 14, and not 25 which it should have been in view of the fact that the sum of the weights (percentages) for the first four groups comes to 75. The figure 14 represents the percentage which expenditure on the *actual* items included in the miscellaneous group bears to the total expenditure of the average working-class family. The weights utilized for the different groups are:

Food	47
Fuel and lighting	7
Clothing	8
House-rent	13
Miscellaneous	14
Total ..	89

The general index is thus constructed with 1933-34 as the base. Since October, 1947 the general index is being shifted to a uniform base period *viz.* August, 1939 by the following purely arithmetical process:—

$$Ia = \frac{Ib \times 100}{Ic}$$

where Ia is the particular index on shifted base of August 1939, Ib the particular index on base 1933-34 and Ic the index for August 1939 on base 1933-34. The index number so converted to August, 1939 is also published in the *Indian Labour Gazette.*

Kanpur Working-Class Cost of Living Index.

The Bureau of Economic Intelligence, U.P. conducted in 1938-39, through the Labour Office, an inquiry into family budgets of mill workers at Kanpur and about 1,500 budgets were collected. Before all these could be compiled, the need of preparing a cost of living index became imperative as questions relating to dearness allowance cropped up as a result of condi-

tions created by World War II. It was, therefore, decided to prepare an interim index on the basis of 300 budgets, which had already been compiled. Thus started the present series of interim cost of living index numbers for mill workers at Kanpur.

The items included in the index are divided into the usual 5 groups. Expenditure on house-hold requisites and miscellaneous items not included in these groups, representing 31% of the total expenditure, has been left out for the purpose of the index.

The weights have been worked out on the basis of percentage expenditure on different items. In addition to the weight for actual expenditure, the weight for wheat-flour was alloted to wheat; for *Jowar* and *Bajra* to *Birra* (mixture of gram and barley); for gram flour to gram; for 'other pulses' to *Arhar Dal;* for beef, fish and eggs to meat; for *Gur* to sugar; for milk, tea, sweets and miscellaneous food articles to *Ghee;* for other oils (excluding hair oil and burning oil) to mustard oil; for vegetables and fruits to potato; and for spices to salt. Weights for apparel of men were alloted to *Dhoti* and those for apparel of women were alloted to *Sari.* Weight for house-rent represents the actual expenditure on the item. In the miscellaneous group, only five items representing 6% of the total expenditure were selected. Except for *bidis,* to which the weight for tobacco was alloted, weights for each item represent the actual expenditure on that item. The commodities selected and the weights assigned are as follows:—

Group Weights in Kanpur Working Class C/L Index.

Group	Weights
1. Food	42
2. Fuel & Light	6
3. Clothing	8
4. House Rent	7
5. Miscellaneous	6
Total ..	69

The commodity weights within each group are:—

Commodity Weights in Kanpur Working Class C/L Index.

Commodities	Weights
I. FOOD	
1. Wheat	29
2. *Birra* (*Bejhar*)	6
3. Gram	3
4. Rice	10
5. *Arhar Dal*	8
6. Meat (Mutton)	6
7. Sugar	3
8. *Ghee*	18
9. Mustard Oil	6
10. Potatoes	8
11. Salt	3
Total ..	100
II. FUEL & LIGHT	
12. Firewood	75
13. Kerosene Oil	25
Total ..	100
III. CLOTHING	
14. Dhoties	69
15. Saries	31
Total ..	100
IV. HOUSE RENT	
16. House Rent	100

V. MISCELLANEOUS

17.	Barber	21
18.	*Dhobi*	23
19.	Soap	12
20.	*Supari*	21
21.	*Bidies*	23
	Total ..	100

The index is prepared from retail prices of the selected commodities collected from 10 shops in various labour *basties* at Kanpur. Formerly, these prices were collected on every Sunday; they are now collected on every Saturday, as markets now close on Sundays at Kanpur. House rent, however, does not admit of detailed enquiry every month and was taken to be constant, to start with. In 1943, a departmental inquiry was made and on the basis of its results the index number for house rent was raised to 121 since January, 1944. A further inquiry was made in January, 1948 and as a result of the information obtained the house rent index was raised to 197 since February, 1948. Prices during August, 1939 have been taken as base prices.

For computing the index, the weekly prices are averaged and the average used as the month's price. The general index as well as the group indices are weighted arithmetic means, like those of the Bombay Working-Class Cost of Living Index Number.

Labour Bureau All-India Working Class C/L Index.

We have seen that the Labour Bureau regularly publishes cost of living indices for 16 centres on more or less uniform lines with a common base, 1944=100 excepting the series for Mercara. The indices compiled by the State Governments are computed on varying base-periods, and even the methods of their compilation are not uniform. However, the series, as they are, are being utilised for various purposes connected with the trend in the movement of prices in items of family consumption in the areas to which they relate.

The existence of a number of series of cost of living index numbers for the various centres in the country with their

diverse trends is more bewildering than helpful. The need for a unified index in the context of all-India was, therefore, keenly felt. It will, no doubt, be fair to appreciate the argument that a single All-India Cost of Living Index has hardly any real significance considering the large variations in family consumption habits and price trends in the various parts of the subcontinent. Notwithstanding this serious limitation, there are problems of an all-India nature which for their consideration require a cost of living index number not exclusively identified with any particular part of the country. To meet this need the question of combining all the existing working-class cost of living index numbers into an all-India index was examined by the Labour Bureau, and they issued, for the first time in December 1952, the All-India Average Working Class Cost of Living Index Number, with the calendar year 1944 as the base. The selection of this base does not involve any adjustment in the case of Labour Bureau series of working-class cost of living index numbers as they are already compiled with base 1944. The States series, which are being constructed on different base periods, have, for the purpose of compiling the All-India Series, been computed on base 1944=100 by simple arithmetical conversion.

Twenty four centres have been selected for the all-India series, *vide* col. (2) of Statement I on next page. Of these, 15 centres are those for which the Labour Bureau compiles the working-class cost of living index numbers and 9 are those for which the State Governments compile similar index numbers.

The all-India index has been compiled by working out a weighted average of the final indices of each component series. For States for which more centres than one have been chosen, the indices for the centres within a State are averaged first to yield a State index, and all the State Indices in their turn are then averaged to yield the All-India Average Index.

As for the weights, it was not possible to ascertain the factory-employment to which a particular component series caters in practice and, as such, weights for the centres within a State are based on the factory-employment at those centres only. The factory-employment is measured by the total number of workers employed in 1944 in the factories registered under the Factories Act, 1934. The number of workers was, of course, corrected for the effects of Partition. These figures for

the various centres and States are given in columns (3) and (4) of Statement I. Column (5) shows the percentage of employment in a selected centre to the total factory employment in the State. It will be seen that the coverage is not satisfactory in the case of Assam, Madras, Orissa, Punjab and West Bengal. The reasons for this state of affairs are the scattered nature of industrial activity in States like Assam and Orissa and the absence of proper series of working-class cost of living index numbers for some of the industrially important centres in the other States.

STATEMENT I.

List of Centres included in the Interim Series of All-India Average Working Class Cost of Living Index Number and their Weights.

State	Centre	Factory Employment in 1944 in the		Percentage coverage of the Centres in the State 4 ÷ 3.	Blown up employment weight used in computation.
		State	Centre		
(1)	(2)	(3)	(4)	(5)	(6)
Part 'A' States					
Assam		53,411			
	Gauhati		754	1.4	22,869
	Silchar		46	0.1	1,395
	Tinsukia		961	1.8	29,147
				3.3	53,411
Bihar		129,247			
	Jamshedpur		40,917	31.7	88,337
	Dehri-on-Sone.		4,336	3.4	9,361
	Monghyr		14,613	11.3	31,549
				46.4	129,247
Bombay		736,464			
	Bombay City		368,035	50.0	506,532
	Ahmedabad		135,297	18.4	186,211
	Sholapur		29,083	3.9	40,027
	Jalgaon		2,684	0 4	3,694
				72.7	736,464

(1)	(2)	(3)	(4)	(5)	(6)
Madhya Pradesh		100,850			
	Jabbalpur		23,685	35.5	48,646
	Akola		5,083	5.1	10,461
	Nagpur		20,324	20.2	41,743
				48.8	100,850
Madras		265,602			
	Madras City		35,380	13.3	265,602
Orissa		6,262			
	Cuttack		1,151	18.4	5,485
	Berhampur		163	2.6	777
				21.0	6,262
Punjab		43,442			
	Ludhiana		3,389	7.8	43,442
Uttar Pradesh		278,238			
	Kanpur		109,545	39.4	278,238
West Bengal		678,591			
	Calcutta		20,014	2.9	83,587
	Howrah		133,758	19.7	558,628
	Kharagpur		8,710	11.3	36,376
				23.9	678,591
Part 'B' States					
Mysore		62,811			
	Bangalore		31,220*	49.7	62,811
Part 'C' States					
Ajmer		15,829			
	Ajmer City		8,498	53.7	15,829
Delhi		27,702			
	Delhi City		25,306	91.4	27,702

The number of centres in a State available for inclusion in the all-India index is largely fortuitous and has little or no relation to its industrial importance. In using centre-weights only, an exaggerated significance may be given to the movements in the index numbers for a State which has a greater number of included centres. By first working out a weighted index number for a State and then combining State indices into an All-India Index by weighting according to the volume of industrial employment in the State, however, the movements in the index numbers for a State are given appropriate importance. All the same, the calculation of a weighted State-Index is

* Estimated.

open to theoretical objections, but it was not possible for the Labour Bureau to do anything better than combine the indices for the available centres on the basis of blown-up weights. This assumes that the available groups of centres in a State are representative of the industrial section of the whole State as regards consumption habits and price trends.

A weighted average of all the component series of index numbers (reduced to base 1944=100) for a month is thus calculated to yield the All-India Average Index, the weights being the blown-up factory-employment at each centre. The 'blown-up factory-employment' shown in col. 6 of Statement I is arrived at by distributing total State-employment-figure over the centres selected within the State in proportion to the factory-employment at each centre.

As an additional point of detail it may be stated that the cost of living index number for Nagpur, as published by the State Government, had to be first adjusted so as to include the House-rent Group with the rent-index kept at 100, as this group is not included in the figures published by the State.

Statement II gives the interim series of All-India Average Working Class Cost of Living Index Number compiled in the above manner:

STATEMENT II.

Interim Series of All-India Average Working Class Cost of Living Index Number.

(Base: 1944=100)

Year Month	Index Number
1945 (Average)	100
1946 (Average)	106
1947 (Average)	120
1948 (Average)	134
1949 (Average)	138
1950 (Average)	138
1951 (Average)	144
1952 (Average)	141
Jan. '53	139

The above table gives an almost continuously rising tendency for the all-India working-class average cost of living since 1945; in 1952 there has been some fall.

The above series is an interim one. There is considerable scope for revision and improvement. The Labour Bureau are, therefore, keeping the basis of the index under review with a view to effecting improvements from time to time. The most satisfactory procedure would, of course, be to conduct fresh family budget enquiries on a well-designed basis, but that would necessarily take time. Meanwhile, as a first step in the direction of compiling an all-India index, the Labour Bureau have proceeded on the basis of combining the existing series. This step is welcome so far as it goes. The series, as it is, includes all Part 'A' States. As for Part 'B' States, only Mysore has an index with pre-1944 year as base and has been included in the all-India average index; Hyderabad and Travancore-Cochin, no doubt, have working-class cost of living indices but their base period refers to post-1944 year and, therefore, these States could not be represented in the all-India average index; other Part 'B' States do not have such indices. Among Part 'C' States also, only Delhi and Ajmer have such indices*. Thus the coverage of the all-India average index is as wide as it could be under the existing conditions.

General Remarks.

The advance made in recent years in India in the matter of construction and presentation of cost of living and retail price index numbers needs to be emphasized. Because of the unsatisfactory character of the retail price index numbers included in the *Index Numbers of Indian Prices* and of the cost of living indices compiled in various centres of India, there has been an evident case for constructing retail price index series in a scientific manner and for compiling cost of living index numbers on a uniform basis. The Rau Court of Enquiry, which was appointed in August, 1940 under the Trade Disputes Act, 1929, to investigate into the dispute regarding Dearness Allowance on the G. I. P. Railway, made the following observations in para 111 of its report:—

* An index number for Bhopal has since been added, but with a post 1944 year as base.

> "None of the cost of living index figures at present available are entirely satisfactory........ The first requisite for any satisfactory revision of the allowances that we have recommended is the preparation of up-to-date cost of living index figures for three distinct classes of areas, city, urban and rural...... We would accordingly recommend that the question of preparing and maintaining such figures for the purposes of the Central Government be considered by the Government of India."

Acting on this suggestion the Government of India formulated a centrally controlled scheme for the preparation and maintenance of cost of living index numbers in selected centres, a brief outline of which was circulated to State Governments in October 1941, to which most of them gave a very encouraging response. The Third Conference of Labour Ministers held at Delhi in January 1942 concluded that it was advisable to ensure uniformity of technique in the compilation of cost of living index numbers in the various States. This was followed by the setting up of an Office of the Director, Cost of Living Index Scheme, by the Government to compile cost of living indices in selected centres on a uniform basis. This Office is now known as the Labour Bureau. The Government also proceeded concurrently with a scheme for the compilation of retail price indices.

Thus, the Government of India are working on three distinct schemes:—

1. The Main Cost of Living Index Number Scheme.
2. Retail Price Index Number Scheme, Urban Centres.
3. Retail Price Index Number Scheme, Rural Centres.

Under Scheme 1, the Labour Bureau are now issuing working-class cost of living index numbers for 15 centres, and plantation-workers' cost of living index for Mercara. The work being done under the other two Schemes has been mentioned in the Section relating to Index Numbers of Prices.

We have also seen that the Government are now compiling and issuing an All-India Average Working-Class Cost of Living Index, and that the working-class cost of living indices, compiled by the respective State Governments are being published after conversion to a uniform base, August 1939. to facilitate

inter-State or inter-Centre comparison. The Government have also now started compiling and publishing Index Numbers of Earnings of Factory Workers in India. The State Governments are also compiling index numbers for certain classes of workers other than industrial workers, such as middle-class families, Government employees, etc. The West Bengal Government, for instance, compile a Calcutta Middle-class Cost of Living Index Number, and Madras Government compile Cost of Living Index Numbers for Low-Paid Employees at 8 Centres in the State. The interest in the construction and use of cost of living index numbers has grown so much now, that even the non-official economic journals are constructing and publishing them. Mention in this connection may be made of the Calcutta Middle-Class Cost of Living Index Number compiled and issued by the *Capital,* an economic weekly.

VII. EARNINGS OF FACTORY WORKERS

As an adjunct to the All-India Average Working-Class Cost of Living Index the Labour Bureau took up the compilation of an All-India Index of Earnings of Factory Workers, with a view to enabling compartive studies of trends of earnings and cost of living to be made. They issued the Earnings Index, for the first time, in February, 1953, in the *Indian Labour Gazette.*

Labour Bureau Index of Earnings of Factory Workers in India.

The index is constructed with 1939 as the base; it is compiled only on annual basis; it is computed in respect of (*a*) 'all industries' for each State, (*b*) each industry for all States combined and (*c*) 'all industries' for 'all States'; it is available from 1944 onwards as the data for the earlier years were not sufficiently detailed. The basis of the index is the statistics of earnings of factory workers collected by the Labour Bureau annually, under the provisions of the Payment of Wages Act, 1936, from factories covered by the Act. The data collected by the Bureau cover most of the *perennial* industries, such as, textiles, engineering, minerals and metals, chemicals and dyes, paper and printing, wood, stone and glass, skins and hides, ordnance factories, mints and miscellaneous industries (comprising such industries as rubber and rubber products, rope works, etc.) which account for nearly 5/6ths of the total fac-

tory employment in the country. Figures relating to railway workshops, covered separately under Railways, are not included in these data. Also not included are the industries falling under the groups "food, drink and tobacco" and "gins and presses" (consisting mainly of seasonal factories, such as, sugar factories, tea factories and cotton ginning and baling factories which work for less than six months in a year) as the information relating to annual earnings of workers employed in these factories would not be comparable with that relating to workers in perennial factories.

As the data are collected under the provisions of the Payment of Wages Act, their scope is determined broadly in terms of the definitions, etc. of that Act. The Act covers factories subject to the Factories Act. Until 1948, the Factories Act covered factories employing twenty workers or more and working with the aid of power. Under the Factories Act, 1948, which came into force in 1949, the scope of the definition of "factory" has been enlarged so to include all factories working with the aid of power and employing ten workers or more, as also those factories working without the aid of power but employing twenty workers or more. During the year 1949, however, not many new factories were registered and there has, therefore, been no appreciable increase in the coverage of the statistics. The provisions of the Payment of Wages Act apply to the employees, including salaried personnel, whose monthly incomes do not exceed Rs. 200. The scope of the statistics used in the index is thus limited to employees drawing a pay of less than Rs. 200 per month.

The term "wages" as defined in the Act includes all remuneration capable of being expressed in terms of money which would, if the terms of the contract of employment were fulfilled, be payable and includes any bonus or other additional remuneration of the nature aforesaid, but does not include the value of any housing accommodation, supply of light, water, medical attendance or other amenities provided by the employers, contributions paid by the employer to any pension fund, or provident fund, etc. The figures of average earnings worked out on the basis of these statistics are, therefore, subject to these limitations. The All-India Average Working-Class Cost of Living Index Number is compiled on base 1944=100. For comparative purposes, it would have been convenient to

compile the Index of Earnings also on the same base. However, so far as this index is concerned, the year 1944 was an abnormal year as it was one of the peak years of the War period, and as certain industries, like the ordnance factories, had gained considerable importance during this period while certain others, which were not so essential from the point of view of the War effort, had relatively lost ground. The distribution of employment as between the various industries during 1944 or afterwards was, therefore, not normal. The year 1939 was considered suitable for adoption as base period.

The formula used for the construction of industry- or State- index is given by

$$I = \frac{100 \times \Sigma(E_o \times e_{eo})}{\Sigma(E_o)}$$

where I is the required index, E_o represents the total earnings in the base year, e_o and e stand for the average earnings in the base and current years respectively and the summation sign extends to 'all industries' or 'all States', as the case may be. Since E_o is equal to $n_o \times e_o$, where n_o is the number of workers employed during the base year, the index can also be expressed as

$$I = \frac{100 \times \Sigma(n_o e)}{\Sigma(n_o e_o)}$$

Suitable adjustments have been made, as far as practicable, in respect of the data for the partitioned States of Bengal, Punjab and Assam for the period prior to Partition, to arrive at Indian Union figures. Similarly, in cases where information was not available for some of the intermediate years, suitable adjustments had to be made in the divisor while computing the relevant indices. Such cases, however, were very few. In spite of all the limitations of the data pointed out above and the adjustments that had to be made to overcome them, the indices arrived at may be considered good enough to indicate the trends in the earnings of factory workers and to give an idea of the movement of real earnings in comparison with the all-India cost of living. Statements I & II on next page give the indices of earnings for each State and for each industry group, respectively.

STATEMENT I.

Index Numbers of Earnings by States.

(Base: 1939=100)

State	Average annual earnings in 1939	Index Numbers					
		1944	1945	1946	1947	1948	1949
	Rs.						
Assam	263.7	236.2	263.5	282.7	302.0	322.7	388.7
Bihar	415.5	129.3	133.8	130.9	204.9	217.2	227.6
Bombay	370.4	235.4	223.6	220.8	267.3	314.7	336.9
Madhya Pradesh	....	258.8	286.2	250.7	325.7	390.1	437.6
Madras	175.9	192.3	204.0	243.3	337.6	371.8	438.9
Orissa	161.8	229.1	230.0	263.7	300.3	402.9	417.9
Punjab	296.0	197.8	209.8	....	234.1	256.2	345.5
Uttar Pradesh	235.6	216.7	223.7	242.2	284.0	380.9	427.2
West Bengal	248.7	170.0	176.7	193.6	222.8	282.3	333.5
Ajmer	163.7	227.1	234.3	280.9	273.3	341.1	343.2
Delhi	309.4	241.7	242.4	240.3	266.3	322.9	323.6
All States	292.3	205.1	205.7	211.0	257.3	308.9	345.5

STATEMENT II.

Index Numbers of Earnings by Industry.

(Base: 1939=100)

Industry	Average annual earnings in 1939	Index Numbers					
		1944	1945	1946	1947	1948	1949
	Rs.						
Textiles	284.0	220.1	217.4	217.4	269.1	326.0	369.1
Engineering	317.0	177.6	191.3	212.2	220.6	280.4	289.1
Minerals & Metals	435.1	136.3	141.0	135.2	212.0	215.1	237.8
Chemicals & Dyes	220.8	227.7	210.2	230.4	288.9	339.0	357.7
Paper & Printing	335.0	139.0	165.0	188.0	215.9	247.7	270.4
Wood, Stone & Glass	193.5	203.4	238.1	244.1	281.1	317.2	340.9
Skins & Hides	286.0	181.6	177.8	188.6	207.9	285.5	312.4
Ordnance factories	362.2	175.0	200.2	221.0	225.8	284.8	351.9
Mints	367.5	200.9	193.7	233.3	292.5	379.3	391.6
Miscellaneous	250.7	202.6	188.1	243.4	255.7	316.5	346.2
All Industries	292.3	205.1	205.7	211.0	257.3	308.9	345.5

We see from the two statements that the general index (covering all States and all industries) rose by about 105 points during the five years ending 1944 and was slightly higher during the next year. Following the end of the World War II in 1945, there was widespread agitation for higher wages among industrial workers. With the achievement of Independence in 1947 and the installation of popular Governments at the Centre and in the States, the demands of the workers for higher wages began to be considered more sympathetically than before, and wages were revised upwards partly through the intervention of Government conciliators, adjudicators, tribunals and courts. In many major industries annual bonuses have since become a regular feature, adding substantially to the earnings of the workers. The effects of all these developments are noticeable in the Index of Earnings, which recorded a remarkable rise of 46 points in 1947, followed by another 52 points in 1948 and 37 points in 1949. As a result of these increases, the index for 'all States' in 1949 stood at nearly 3½ times its level in 1939. Compared to this rise in the general index, the increase in the State's index was higher in Assam, Madhya Pradesh, Madras, Orissa and Uttar Pradesh. A similar trend is observed in industries index numbers, though it is not so marked as in the case of the regional indices; in 1949 the index numbers of earnings were higher than the general index in Textiles, Chemicals and Dyes, Ordnance factories, Mints and Other Miscellaneous Industries.

All-India Earnings and Cost of Living Indices Compared.

A comparison of the indices of earnings with those of cost of living would indicate the trend in the *real* earnings of workers. For purposes of comparison, the Labour Bureau has shifted the base of the Index of Earnings to 1944 and worked out a new series. The series thus worked out and the all-India cost of living index series with the same base are given in Statement III on the next page. The third row in the Statement gives the indices of *real* earnings, computed by deflating the Earnings index for each year by the Cost of Living index for that year.

profits and that in the case of jute, tea, paper, iron and steel and cement the profits were at their peak in 1950 during the period under review.

IX. FOREIGN TRADE

The Department of Commercial Intelligence and Statistics has, for many years past, been constructing index numbers of (*a*) average unit declared values (prices) of imported and exported merchandise and (*b*) quantum of foreign (sea and air-borne) trade of India, and issuing them through the *Accounts relating to the Foreign (Sea, Air and Land) Trade and Navigation of India,* from where they are reproduced in several official and non-official journals. For some years past the Reserve Bank of India has also been compiling similar index numbers of (*a*) unit value and (*b*) quantum and issuing them through the *Reserve Bank of India Bulletin.* The base year of both the series is 1948-49.

D.C.I. & S. Index of Unit Value and Quantum.

The D.C.I. & S. series of unit value and quantum index numbers are compiled and published on monthly and annual basis. The formula used in constructing the unit value index is the aggregative weighted average formula (also called, Paasche type, *cf.* formula (2), p. 246, Chapter XI), current quantity figures being used as weights. In the quantum index, the formula used is of Laspeyres type (*cf.* formula (1), p. 246, Chapter XI), that is, base-period value figures are used as weights. The quantum index is, in fact, not directly calculated, but is derived from the value ratio and the price index referred to in the unit value index. The calculations for the D.C.I. & S. series are based on selected items relating to foreign trade by sea and air, the items not directly used in constructing the index being assumed to be subject to the same price changes as certain other groups of related items. This method of adjustment is adopted in certain foreign countries too. The over-all indices of the two series of D.C.I. & S. are published along with three group-indices.

Reserve Bank's Index of Unit Value & Quantum.

These series are compiled and issued on monthly, quarterly and annual basis. The formulae used for both the unit value

and quantum indices are of Laspeyre's type, that is, base-period weights are used. The calculations in this case are also based on selected items, but with two differences as compared with the D.C.I. & S. series: firstly, the foreign trade by land is also included along with that by sea and air; and, secondly, no adjustment is made for items not selected. The first difference constitutes an improvement over the D.C.I. & S. series. The second is a drawback; the proportion of trade covered by the index varies from month to month which may lead to wrong conclusions. The over-all indices of the two series of R.B.I. are published along with 3 group indices, and 16 sub-group indices for exports and 20 sub-group indices for imports.

D.C.I. & S. and R.B.I. Series Compared.

The table below compares the two series.

Quantity and Price Indices of Imports and Exports
D.C.I. & S. and R.B.I. Series.

(Base: 1948-49=100)

Period	QUANTITY INDEX				PRICE INDEX			
	Imports		Exports		Imports		Exports	
	DCIS	RBI	DCIS	RBI	DCIS	RBI	DCIS	RBI
Annual								
1951-52	118	108	98.1	89	130.2	147	168.9	178
1952-53	90	74	109	94	125.8	128	118.7	116
Monthly								
Mar. '52	141	118	99	83	133.8	154	150.0	148
Mar. '53	85	60	115	101	114.2	123	110.1	103

It will be seen from the above table that while the variations over time in the two series run in the same direction, the absolute values of the same category of index numbers in the two series are not identical. This conflict arises mainly from dif-

ferences in coverage and method of construction of the two series.

Both the series are useful in their respective spheres. The D.C.I. & S. series may be used as a measure of long term movements in prices, while the R.B.I. series as that of price changes over short periods. The D.C.I. & S. series is suitable for deflating current value totals of imports and exports to the price level of the base year; it is well suited for the measurement of terms of trade.

In order, however, that the two series may serve better their respective purposes, the R.B.I. series should provide for making adjustments for items not directly used and the coverage of the D.C.I. & S. series should be extended so as to include foreign trade by land also.

CONCLUSION

General Appraisal.

It should be clear from the foregoing that index numbers are now available on a variety of economic subjects in India. The quality of these index numbers and the extent to which they correctly present the trends of the phenomena to which they relate would naturally depend on the quality of the data which form their basis. The drawbacks and limitations to which the different primary statistics of production, prices, foreign trade, etc. are subject have already been discussed in Chapter XXII; the more important limitations of the data entering the different index numbers have also been mentioned in the relevant sections of this Chapter. What is stressed here is that before attempting to draw conclusions from an index number series, one must devote thought to the limitations of the data on which the given series is based and to the adjustments that may have been made by the compiler of the series to remove any bias resulting from such limitations. It should also be the duty of those collecting and compiling the primary statistics to see that as many drawbacks or limitations of the basic statistics are done away with as possible.

Along with attempts at improving the quality of the basic data, it is necessary to keep the method of construction of an index number under constant review so as to utilise improved

methodology from time to time. No one method of construction is good enough for all index numbers; for the same index number, no one method is good enough for all times. At any given time, the choice of the method is dependent on the nature of basic data available, on the kind of material available for determining weights, and on the purpose which the index is intended to serve. As these guiding factors change or improve, the methodology should come up for a review. This explains, for instance, why the Bombay Working Class Cost of Living Index first published in 1921 was based on the aggregate consumption method and later, on the collection of weights from the family budget enquiries, the method was revised; or, why the Reserve Bank of India revised, with effect from April 1953, its Security Prices Index; or, why the Economic Adviser gave up the sensitive wholesale price index and initiated the general purpose series. The general line adopted in this matter is to make a beginning with as perfect an index as the available data permit, and to revise it when more information is gathered. The official index of industrial production, for that reason, is still 'interim' and the same is the case with the all-India average index of working class cost of living. If, however, the quality of basic data change materially after an index number series has run for some time, it would be better to suspend or discontinue the series rather than present a misleading picture by continuing it on the basis of incomparable data. It will be well to recall the circumstances under which the Ministry of Food and Agriculture index numbers of Indian Agricultural Production have been suspended, and the several 19th century indices discontinued.

While reviewing the method of construction of an index number series several points call for attention. If the scope of primary statistics has widened so that data are now being collected on a larger number of items than before and/or data on the same items are being collected from a larger area than before, the list of items entering the index should be expanded. If necessary, the grouping of individual commodities into sub-groups and groups should be revised. If the nature of information available for determination of weights has improved, weights should be revised. If the method of averaging requires change on scientific grounds, if the need for continuing the chain-base system does not any longer exist, if the presentation

of the index numbers needs improvement—in all such cases the needful should be done.

An important point may be emphasised in this connection. We have seen that in a weighted index number, two types of base years are required to serve two distinct purposes: One is the comparison base and the other the weight base. In theory, there is nothing wrong in keeping the comparison base different from the base year for weights. For instance, in the Reserve Bank's old series of Security Price Index Numbers, while the comparison base was 1938 the weight base was 1946. In practice, however, it gives a clearer meaning to the index if the two types of bases are identical.

Common Base Period for Official Index Numbers.

We have seen that not all official index numbers in India are, at present, being compiled on one common comparison base. On the other hand, to study related economic phenomena, as for instance, production and prices, or cost of living and earnings, it would be of definite advantage to have the concerned indices based on a common base period so that a complete picture of relative movements in the different phenomena would be brought to light, just at a glance. One way of doing it would be to shift the base of one series to that of another by mere arithmetical conversion. The conversion of States' cost of living index numbers to August, 1939, before their publication in the *Indian Labour Gazette*, is a case in point. This method is, however, rough and ready, and is not free from defect in certain cases. The other and the more scientific method would be to reconstruct the entire series or set of series on one common base year. Here too, a question of principle arises: If index number theory requires that the base year for each index number series should be a *normal* year in respect of the particular phenomenon it measures, and if two or more related phenomena do not have the same year as a *normal* year, then how to select a common base year for the concerned series? The answer, naturally, is that a compromise should be found out, that a *period* of more than a year or two years should be selected within which the base *years* or base *months* of the different series should be located so that even though the base years or base months may not be common for the different series, they

would be very close to one another and would not materially affect the comparison.

The weight of international practice is largely in favour of a common base period for all, or, at any rate, related, index numbers. In U.S.A., Canada, France and South Africa common base periods have been adopted for their respective economic indicators. In India, too, whenever revision of any existing series is taken up, or a new series is constructed, this international practice should be followed. This has become all the more necessary now that a new era of economic planning has opened in the country and the lay public as well as the research workers would be naturally interested in measuring the progress in one phenomenon in relation to that in others, and this study can be facilitated only if a common base period is adopted for the different economic indicators. It cannot also be denied that many index numbers in India now require revision both in respect of the comparison base and weight base owing, *inter alia*, to events like Partition, Devaluation, Korean War, change-over from controlled economy to comparatively more free economy, etc. Base years such as the pre-War years are now clearly inappropriate; they are far too away from the present while a base period should be as near the period studied as possible. In view of the fact that revision of base years in Indian Index Numbers is now urgent, it would be exceedingly good if the opportunity were utilized for (*a*) having a co-ordinated and integrated plan for a uniform comparison base period for the official index numbers and (*a*) revising the weight base so as to bring it in conformity with the revised comparison base. The latter, no doubt, requires a huge volume of work in respect of index numbers like those of cost of living, where family budget inquiries must be conducted before weights can be determined. The labour and expense involved in conducting such enquiries are, however, worth incurring now.

Lacunae.

The volume and scope of index numbers have been growing in India in recent years, yet there are certain lacunae which need to be urgently filled up. An account of the steady growth of index numbers in India has already been given in this Chapter together with that of improvements effected in the revised

series. Suggestions for further improvements in the existing series have also been made. It is time now that attention should be devoted to the construction of economic indicators like the index numbers of wage rates, retail trade, inland trade, agricultural prosperity, food supply, population, etc. as well as monetary indices. A good deal of basic statistics on these subjects are now available and more could be collected. Agriculture being our mainstay, index numbers of agricultural prosperity should naturally receive the highest priority. These index numbers imply the construction of two series of index numbers: (1) index numbers of prices received by the farmers and (2) index numbers of prices paid by the farmers. The latter include the cost of living as well as the expenses of cultivation. When these two series have been compiled, it is easy to work out the parity between them in order to measure the progress of agricultural prosperity in the country. The construction of these index numbers, no doubt, calls for much spade work to be done, for instance conducting the necessary enquiries in the rural areas to determine the items entering the farmer's cost of living and expenses of cultivation and to estimate the weights. In the interest, however, of providing an instrument for measuring the prosperity of 4/5ths of India, the finances and labour required for producing this instrument should be found out.

EXERCISES

(1) What purpose is served by index numbers? Appraise the principal index numbers published in India.

(M.A., Alld., 1951).

(2) What index numbers of agricultural production in India are available? Name the agencies that compile them and discuss the method of compilation adopted by any *one* of them.

(3) Examine the limitations of the index numbers of agricultural production in India compiled by the Reserve Bank of India and the *Eastern Economist* and give suggestions for their improvement.

(4) Explain the scope and method of construction of the official index of industrial production in India. Mention also how adjustment is made in the index for seasonal variations in the output of sugar industry.

(5) What light does the table below throw on the progress of industrialisation in India?

Interim Index of Industrial Production.

(Base: 1946=100)

Industry		1948	1949	1950	1951	1952	Apr. 1953
General Index	....	108.4	106.3	105.2	117.4	128.5	141.2
Sugar	....	116.5	108.5	105.8	120.8	161.9	459.3
Cement	....	100.7	136.3	169.4	207.2	229.3	224.6
Plywood	....	91.4	81.2	85.4	120.5	154.0	124.1
Cotton Cloth	....	110.5	99.9	93.8	104.3	117.7	130.4
Jute Mfs.	....	100.2	86.9	76.8	80.3	87.4	83.0
Coal	....	103.2	108.9	110.8	118.8	125.4	136.6
Electric fans	....	163.0	162.1	175.3	192.5	177.1	183.9

(6) Explain the utility of harvest price index numbers in an agricultural country like India. How will you compile such an index for India?

(7) Examine the nature, scope and construction of Harvest Price Index Numbers in India. What is their utility?

(8) Explain the difference between index numbers of retail prices and cost of living. Illustrate your answer with the help of such index numbers constructed in India.

(9) How are retail price index numbers available in India constructed? Examine their usefulness in economic discussions.

(10) Explain the scope and method of construction of the revised series of index numbers of security prices in India.

(11) In what respects is the revised series of security price index numbers compiled by the Reserve Bank of India an improvement over the old series? What purposes do these index numbers serve?

(12) What is the objective of an index number? State briefly the relevant conditions for its construction illustrating your answer by reference to the Index Numbers of Prices published by the Government of India.

(B. Com., Bombay, 1936).

(13) Describe any index number in use in India at present for measuring changes in the wholesale price level, and point out its shortcomings.

(M.A., Cal., 1937).

(14) Examine critically the Calcutta wholesale price index prepared by the Director-General of Commercial Intelligence and Statistics. In what respects is the new index prepared by the Economic Adviser to the Government of India an improvement upon the Calcutta index?

(M.A., Patna, 1943).

(15) Describe carefully how you would proceed to construct the cost of living index numbers for the U.P. (for the benefit of industrial labour). Would you allot weights according to 'Fisher's Ideal Method' or 'Family Budget Method'? Give reasons in support of your answer.

(16) Explain clearly how the "Capital" Index of Industrial Activity in India is calculated. How far do you consider it representative?

(B. Com., Alld., 1940).

(17) What statistical material would you utilize for preparing an index of economic activity in India?

How would you collate your data?

(M. Com., Luck., 1942).

Name the statistical publications which you would have to consult.

(M.A., Alld., 1950).

(18) What is the technique of the cost of living index numbers? What is the method of their compilation in India and for what places are they compiled?

(B. Com., Nagpur, 1943).

(19) Examine critically the construction of *either* the Economic Adviser's Index of Wholesale Prices *or* the Working Class Cost of Living Index Number as compiled by the Government of India.

(M.A., Alld., 1952).

(20) Explain the method of construction of *either* the Bombay Working-class Cost of Living Index *or* the Wholesale Price Index issued by the Economic Adviser to the Government

of India, giving information regarding the following points particularly:

(*a*) Base period,
(*b*) Groups of commodities included,
(*c*) Method of weighting and averaging adopted.

(M.A., Raj., 1951).

(21) What index numbers are being compiled and issued by the Labour Bureau in India? Explain in detail the construction of their one of the all-India series.

(22) The Rau Court of Enquiry, 1940, reported that none of the cost of living index numbers then available was entirely satisfactory. What has the Government done since to produce satisfactory index numbers?

(23) Explain and illustrate the utility of index of working class cost of living and earnings of factory workers, and give the method of construction of *either* of the two series available in India.

(24) What are the limitations of the index numbers of earnings of factory workers compiled in India? How are these indices compiled?

(25) What series of index numbers are available in India in respect of quantum and unit value of exports and imports? How do the available series differ in regard to their scope and method of construction? Examine their usefulness.

(26) What distinct purposes are served by the index numbers of foreign trade compiled by the Department of Commercial Intelligence and Statistics and the Reserve Bank of India? What improvements should, in your opinion, be made in the two series so that they may serve their individual purposes better?

(27) Explain the usefulness and method of construction of *either* the index numbers of industrial profits *or* those of industrial activity.

(28) 'No one method of construction is good enough for all index numbers; for the same index number, no one method is good enough for all times'.

Explain and illustrate the above statement with Indian examples.

(29) What considerations will weigh with you at the time of revising an index number series? What importance would you attach to uniformity in comparison base and weight base in this connection?

(30) 'The volume and scope of index numbers have been growing in India in recent years, yet there are certain lacunae which need to be urgently filled up'.—Discuss.

(31) Examine critically the Economic Adviser's Index Number of Wholesale Prices and suggest ways to improve it.

(M. Com., Alld., 1951).

(32) Explain the method of construction of any cost of living index number in India, giving information regarding the following points particularly:

(*i*) authority compiling the index number;
(*ii*) base period of the index;
(*iii*) groups of commodities included;
(*iv*) method of weighting adopted and the manner of arriving at the weights;
(*v*) method of averaging used.

(B. Com., Delhi, 1953).

APPENDIX I A.

SPECIMEN OF A BLANK-FORM

The first round of the Government of India National Sample Survey was conducted during October 1950—March 1951. In this round information was collected on consumption, prices, wages, etc., in the randomly selected villages for the period July-June 1950. Two series of schedules were used in this round: One series was prepared in the Indian Statistical Institute, Calcutta, and the other in the Gokhale Institute of Politics and Economics, Poona. The first series was used in one set of 1189 villages and the other in a second set of 644 villages.

The schedules prepared in the Indian Statistical Institute may be grouped into 4 types :

(1) Village schedules ;

(2) Household schedules for general particulars on demographic and economic conditions ;

(3) Household schedules for detailed information on house-hold enterprise ; and

(4) House-hold schedules for detailed information on consumer expenditure.

The first group included 5 schedules. They were numbered as 0, 0.1, 4, 5 and 6. Part I of Schedule 6 was meant for collecting prices of certain selected commodities in the sample villages and Part II for collecting rates of daily wages of various types of skilled and unskilled workers, particularly in relation to agricultural operations. The commodities covered by Part I included cereals, pulses, oil, milk, vegetables, spices, fuel and tobacco. A specimen of the blank form used for Part I of Schedule 6 is given on the next two pages.

GOVERNMENT OF INDIA NATIONAL SAMPLE SURVEY: 1950-51.

Schedule 6: Fortnightly Prices and Wages Report.

Part I. Retail prices as in month................195....

1		2		3	
1. Range		1. Stratum		1. District	
2. State		2. Substratum		2. Tehsil	
3. Block		3. Subsample		3. Union/firka	
4. Scheme		4. Sample unit I		4. Mouza J. L. No.	

4

1. Place where market serving the village is situated............ distance from centre of village............miles.
2. Date on which prices collected: (1) for cols. 4 and 10.......... (2) for cols. 5 and 11................
3. Local unit of weight (........)=........seers........tolas (in standard seers of 80 tolas each of 180 grains) (b).
4. Local unit of measure (........)=........seers, taking one seer as=volume of one standard seer of water (if friction, use decimals).

5

Commodity	Specification	Local unit	Price per local unit in annas as on		Source (a)
			10th	25th	
1	2	3	4	5	6
1. Rice	medium				
1a. ,,	coarse				
2. Rice flour					
3. Wheat					
4. Wheat flour					
5. White flour (maida)					
6. Millets					
7. Gram					
8. Other cereals					
9. Arhar pulse					
10. Gram pulse					
11. Mung					
12. Kalai					
13. Other pulse					
14. Groundnut					
15 Copra					

PART I—continued from pre-page.

Commodity	Specification	Local Unit	Price per local unit in annas as on		Source (a)
			10th	25th	
1	2	3	4	5	6
16. Coconut oil					
16a. Groundnut oil					
17. Mustard oil					
18. Linseed oil					
19. Cow ghee					
20. Buffalo ghee					
21. Cow milk					
22. Buffalo milk					
23. Butter					
24. Potato					
25. Leafy vegetable					
26. Non-leafy ,,					
27. Fruit		100			
28. Salt					
29. Dry chillies					
30. Onions					
31. Turmeric					
32. Mutton					
33. Fresh fish					
34. Dry fish					
35. Eggs	hen/duck	doz.			
36. Sugar					
37. Gur (molasses)					
38. Coke					
39. Fire wood					
40. Kerosene	white				
41. Green fodder					
42. Grass					
43. Hooka tobacco	common quality				
44. Chewing tobacco	—do—				
45. Biri	—do—	100			
46. Handloom dhuti	—do—	adult size pair			
47. ,, ,, sari	—do—	—do—			
48. ,, lungi	—do—	—do—			

(a) Codes for noting source: from permanent shop (1), from peripatetic shop (2), from grower-seller (3), from other sources (4).

(b) Notes regarding special local weight or measures.

Signature of Investigator.............. Date............

APPENDIX I B.

SPECIMEN OF A QUESTIONNAIRE

The Government of India appointed a Taxation Enquiry Commission in 1953 with the following terms of reference:—

(1) To examine the incidence of Central, State and Local taxation on the various classes of people and in different States ;

(2) To examine the suitability of the present system of taxation—Central, State and Local—with reference to (a) the development programme of the country and the resources required for it, and (b) the objective of reducing inequalities of income and wealth ;

(3) To examine the effects of the structure and level of taxation of income on capital formation and maintenance and development of productive enterprise ;

(4) To examine the use of taxation as a fiscal instrument in dealing with inflationary or deflationary situations ;

(5) To consider other relevant matters ; and

(6) To make recommendations, in particular, with regard to (a) modifications required in the present system of taxation, and (b) fresh avenues of taxation.

The Commission issued a questionnaire on the 17th July, 1953. The questionnaire was divided into 6 Parts as under:—

Part I—The Tax System.
Part II—Direct Taxes.
Part III—Commodity Taxes (Central and State).
Part IV—Agricultural Income-tax, Land Revenue and Irrigation Rates.
Part V—Other Taxes (Central and State).
Part VI—Local Taxation.

Part I of the questionnaire was divided into the following sections:—

(1) General,
(2) Incidence of Taxation,
(3) Taxation and economic development, and

(4) Taxation and inflationary and deflationary situations.

The first 11 questions in Part I fell under the General Section. These questions are reproduced below:—

QUESTIONNAIRE

PART I—THE TAX SYSTEM.

1. What, in your opinion, should be the objectives of a sound tax policy? What should be the place in such a policy of the following objectives:

(a) reduction in inequalities of income and wealth;
(b) encouragement of incentives to work, to save and to invest;
(c) countering of inflationary and deflationary tendencies; and
(d) maintenance of the external balance of the economy?

In what directions would you suggest modification of the Indian tax system having regard to these objectives?

2. What criteria would you suggest for determining the equity of a tax system. How far is it possible to secure it in the Indian tax system?

3. Does the Indian tax system conform adequately to the principle of ability to pay or do you think an extension of this principle is desirable and practicable? If so, in what ways?

4. Differing views have been expressed on the extent to which the taxable capacity of various sections of the community has been used by the Indian tax system. What is your view on the subject?

5. The proportion of tax revenue to national income in India is considered small relatively to several other countries. Do you think this proportion can be raised, and, if so, what tax changes would you suggest for the purpose?

6. Do you think that the relative place of direct and indirect taxes in the Indian tax system is satisfactory?

7. Do you consider that non-tax revenues are likely, in future, to occupy a more important place than hitherto in the

public revenues of India ? If so, please state the sources of non-tax revenue you have in view and indicate to what extent they may be expected to contribute to the public exchequer.

8. Do you agree with the view that receipts from particular taxes should not, as a rule, be funded or ear-marked for specific purposes ?

9. Do you think that it is desirable, under certain conditions, to levy cesses for special purposes ?

10. State undertakings, commercial, industrial, etc., are coming to play an increasingly important part in the economy of the country. Have you, from the point of view of their significance as a source of revenue, actual or potential, any comments or suggestions to make on matters such as the further extension of State undertakings and their policies in regard to pricing, in so far as these may be relevant to tax policy ?

11. Would you suggest that the net surplus earned by State undertakings should accrue to general revenues or be carried to a fund for financing projects of development or be re-invested in the undertakings concerned ?

APPENDIX II.

LOGARITHM.

Logarithm of a given number to the base ten is the power to which the base ten should be raised to equal the given number.

$10000 = 10^4$	∴	Logarithm	of	$10000 = 4$
$1000 = 10^3$	∴	,,	,,	$1000 = 3$
$100 = 10^2$	∴	,,	,,	$100 = 2$
$10 = 10^1$	∴	,,	,,	$10 = 1$
$1 = 10^0$	∴	,,	,,	$1 = 0$
$.1 = 10^{-1}$	∴	,,	,,	$.1 = -1$
$.01 = 10^{-2}$	∴	,,	,,	$.01 = -2$
$.001 = 10^{-3}$	∴	,,	,,	$.001 = -3$
$.0001 = 10^{-4}$	∴	,,	,,	$.0001 = -4$

From the above it will be seen that the log. of 1 is 0 and log. of 10 is 1. Therefore, log. of any number between 1 and 10 would be greater than zero but less than 1. That is, it would be equal to 0 + a fraction. Similarly, log. of any number between 10 and 100 is 1 + a fraction. Thus a logarithm may consist of two parts : the whole number, and the fraction. The whole number part, *e.g.*, 0 or 1 in the above instances, is termed **characteristic** and the fractional part is termed **mantissa.**

To determine the characteristic of any given number, the following two rules should be noted :

(1) If the given number is greater than one, the characteristic is always positive, and is obtained by the formula $(n-1)$, where n stands for the number of significant digits before the decimal point.

(2) If the given number is less than one, the characteristic is always negative, and is obtained by the formula $(N+1)$, where N stands for the number of zeroes after the decimal point but before any significant digit. The minus sign of the negative characteristic is written at the top of the characteristic and not prefixed to it. Thus, the characteristic of minus two is written as $\bar{2}$ and not as -2.

In accordance with the above two rules characteristics of a few numbers are given below :

Number	Value of n (rule 1)	Characteristic
4539	4	3
453.9	3	2
45.39	2	1
4.539	1	0

Number	Value of N (rule 2)	Characteristic
.4539	0	$\bar{1}$
.04539	1	$\bar{2}$
.004539	2	$\bar{3}$
.0004539	3	$\bar{4}$

Characteristic for any number can be similarly calculated.

For calculating mantissa for different numbers, Logarithmic Table has to be consulted. Logarithmic Table giving the mantissa for any number having less than four digits is given at the end of this Appendix. Mantissa of the required number should be read out from this table irrespective of the position of the decimal. If the given number is composed of four or more digits, it must first be approximated to 3 digits. Then its mantissa should be read out from the table. If we are to find the logarithm of 4539, we first approximate it to three digits. The approximation is 454. The characteristic of 4539 is 3, and the mantissa of 454 is .6571. Therefore, logarithm of 4539 is 3.6571.

Two facts about mantissa should be clearly noted. Firstly, it is always positive irrespective of the fact whether the characteristic is negative or positive. Secondly, mantissa is not affected by the position of the decimal point in the number. The mantissa of 454, 45.4, 4.54, .454, .0454 or 45400000 is the same ; only the characteristic will differ. Thus, to find the logarithms of the numbers whose characteristics are given above, .6571 would be attached to the characteristics already ascertained. The logarithm of 4.539. for instance, would be 0.6571, and that of .004539 will be $\bar{3}.6571$. Logarithm of any given number can be similarly calculated. In adding up a series of logarithms in which some characteristics are negative and some positive, the mantissa of all the logs, should be taken

as positive, and characteristics should be treated according to their algebraic signs.

Antilogarithm.

Antilog. of a given number is the number, logarithm of which is the given number. An Antilog. Table is also given at the end of this Appendix. With its help antilog. of any number can be determined. The mantissa of the given number will give the different digits of the required number and characteristic will enable us to locate the decimal point in it. Suppose we want to find the antilog. of 2.1563. We approximate the mantissa, .1563, to three digits and it comes to .156. The antilog. of .156 is 1432. As the characteristic is + 2, according to rule 1 there must be three significant digits before the decimal point. Therefore, the required number is 143.2. Similarly, to find the antilog of $\bar{2}.1563$, we ascertain the antilog. of .156 which is 1432, as in the former case. But since the characteristic is minus, according to rule 2 there must be one zero after the decimal point, so that the required number is .01432. Antilog. of any number can be similarly calculated.

Use of Logarithms.

The logarithm of the product of any two numbers is the sum of the logs. of the two numbers taken separately and, therefore, when two or more numbers are to be multiplied the logarithm of each number is found out and added. The antilog. of the sum is the required product.

$$\text{Log. } (a \times b) = \text{Log. } a + \text{Log. } b$$

$$\therefore \; a \times b = \text{Antilog. [Log. } a + \text{Log. } b]$$

The logarithm of any number a divided by b is the difference of the logs. of a and b. Therefore, when one number is divided by another, the logs. of both the numbers are found out and the antilog of the difference between the two logs. given the desired quotient.

$$\text{Log. } \frac{a}{b} = \text{Log. } a - \text{Log } b.$$

$$\therefore \; \frac{a}{b} = \text{Antilog. [Log. } a - \text{Log. } b]$$

The logarithm of any number raised to n^{th} power is n times the log. of that number. Therefore, when any given number is raised to any power, the logarithm of the given number is found and multiplied by the power to which the number has been raised. The antilog. of the product gives the value of the given number raised to the desired power.

$$\text{Log. } a^n = n \text{ Log. } a$$

$$a^n = \text{Antilog. } [n \text{ Log. } a]$$

The log. of a given number to the nth root is equal to the log. of that number divided by n. Therefore, when the value of any given number to any given root is desired to be obtained, the log. of the given number is found out. This log. is divided by the given root to which the given number is to be reduced. The antilog. of the quotient gives the value of the given number reduced to the desired root.

$$\text{Log. } \sqrt[n]{a} = \frac{1}{n} \text{Log. } a$$

$$\therefore \sqrt[n]{a} = \text{Antilog.} \left[\frac{\text{Log. } a}{n} \right]$$

MATHEMATICAL TABLES

INSTRUCTIONS.

Table of Logarithms—This table gives the *mantissa* of figures. To find the *mantissa* of any given number from the table, the number should be approximated to three digits. *Mantissa* of a number is the same regardless of the position of the decimal point in it.

Table of Antilogarithms—This table gives the antilogarithms of the *mantissa* portion of any given logarithm. The position of the decimal point in the required figure should be determined on the basis of the characteristic of the given logarithm.

Table of Squares—In this table upto 316 one zero, and from 317 onwards two zeroes, are omitted in each square. If in the given figure the decimal point moves by *one* digit to the left, then the decimal point moves by *two* digits to the left in the square.

Table of Square Roots—This table gives two square roots for each number. For *odd digits* in the given number, the upper figure, and for *even* digits the lower figure should be taken. If in the given number the decimal point moves by *two* digits to the left, then the decimal point moves by *one* digit to the left in the square root.

Table of Reciprocals—If in the given number the decimal point moves by *one* digit to the right, then the decimal point moves by *one* digit to the left in the reciprocal.

I

ILLUSTRATIONS

A few illustrations of arriving at logs. etc. from the Mathematical Tables are given below :

Logs., anti-logs. and reciprocals.

(1) Number	(2) Log. of col. (1)	(3) Anti-log. of col. (2)	(4) Reciprocal of col. (1)
0.009	3.9542	0.009	111.1
0.09	2.9542	0.09	11.11
0.9	1.9542	0.9	1.111
9.0	0.9542	9.0	0.1111
85.9	1.9340	85.9	0.01164
859.0	2.9340	859.0	0.001164
1230.0	3.0899	1230.0	0.000813

Squares and square-roots.

(1) Number	(2) Square of col. (1)	(3) Square-root of col. (1)	(4) Square-root of col. (2)
0.009	0.000081	0.09487	0.009
0.09	0.0081	0.3	0.09
0.9	0.81	0.9487	0.9
9.0	81.0	3.0	9.0
85.9	7379	9.269	85.9
859.0	737900	29.31	859.0
1230.0	1513000	35.07	1230.0
316.0	99860	17.78	316.0
317.0	100500	17.80	317.0

LOGARITHMS

	0	1	2	3	4	5	6	7	8	9
10	·0000	0043	0086	0128	0170	0212	0253	0294	0334	0374
11	·0414	0453	0492	0531	0569	0607	0645	0682	0719	0755
12	·0792	0828	0864	0899	0934	0969	1004	1038	1072	1106
13	·1139	1173	1206	1239	1271	1303	1335	1367	1399	1430
14	·1461	1492	1523	1553	1584	1614	1644	1673	1703	1732
15	·1761	1790	1818	1847	1875	1903	1931	1959	1987	2014
16	·2041	2068	2095	2122	2148	2175	2201	2227	2253	2279
17	·2304	2330	2355	2380	2405	2430	2455	2480	2504	2529
18	·2553	2577	2601	2625	2648	2672	2695	2718	2742	2765
19	·2788	2810	2833	2856	2878	2900	2923	2945	2967	2989
20	·3010	3032	3054	3075	3096	3118	3139	3160	3181	3201
21	·3222	3243	3263	3284	3304	3324	3345	3365	3385	3404
22	·3424	3444	3464	3483	3502	3522	3541	3560	3579	3598
23	·3617	3636	3655	3674	3692	3711	3729	3747	3766	3784
24	·3802	3820	3838	3856	3874	3892	3909	3927	3945	3962
25	·3979	3997	4014	4031	4048	4065	4082	4099	4116	4133
26	·4150	4166	4183	4200	4216	4232	4249	4265	4281	4298
27	·4314	4330	4346	4362	4378	4393	4409	4425	4440	4456
28	·4472	4487	4502	4518	4533	4548	4564	4579	4594	4609
29	·4624	4639	4654	4669	4683	4698	4713	4728	4742	4757
30	·4771	4786	4800	4814	4829	4843	4857	4871	4886	4900
31	·4914	4928	4942	4955	4969	4983	4997	5011	5024	5038
32	·5051	5065	5079	5092	5105	5119	5132	5145	5159	5172
33	·5185	5198	5211	5224	5237	5250	5263	5276	5289	5302
34	·5315	5328	5340	5353	5366	5378	5391	5403	5416	5428
35	·5441	5453	5465	5478	5490	5502	5514	5527	5539	5551
36	·5563	5575	5587	5599	5611	5623	5635	5647	5658	5670
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786
38	·5798	5809	5821	5832	5843	5855	5866	5877	5888	5899
39	·5911	5922	5933	5944	5955	5966	5977	5988	5999	6010
40	·6021	6031	6042	6053	6064	6075	6085	6096	6107	6117
41	·6128	6138	6149	6160	6170	6180	6191	6201	6212	6222

LOGARITHMS

	0	1	2	3	4	5	6	7	8	9
42	·6232	6243	6253	6263	6274	6284	6294	6304	6314	6325
43	·6335	6345	6355	6365	6375	6385	6395	6405	6415	6425
44	·6435	6444	6454	6464	6474	6484	6493	6503	6513	6522
45	·6532	6542	6551	6561	6571	6580	6590	6599	6609	6618
46	·6628	6637	6646	6656	6665	6675	6684	6693	6702	6712
47	·6721	6730	6739	6749	6758	6767	6776	6785	6794	6803
48	·6812	6821	6830	6839	6848	6857	6866	6875	6884	6893
49	·6902	6911	6920	6928	6937	6946	6955	6964	6972	6981
50	·6990	6998	7007	7016	7024	7033	7042	7050	7059	7067
51	·7076	7084	7093	7101	7110	7118	7126	7135	7143	7152
52	·7160	7168	7177	7185	7193	7202	7210	7218	7226	7235
53	·7243	7251	7259	7267	7275	7284	7292	7300	7308	7316
54	·7324	7332	7340	7348	7356	7364	7372	7380	7388	7396
55	·7404	7412	7419	7427	7435	7443	7451	7459	7466	7474
56	·7482	7490	7497	7505	7513	7520	7528	7536	7543	7551
57	·7559	7566	7574	7582	7589	7597	7604	7612	7619	7627
58	·7634	7642	7649	7657	7664	7672	7679	7686	7694	7701
59	·7709	7716	7723	7731	7738	7745	7752	7760	7767	7774
60	·7782	7789	7796	7803	7810	7818	7825	7832	7839	7846
61	·7853	7860	7868	7875	7882	7889	7896	7903	7910	7917
62	·7924	7931	7938	7945	7952	7959	7966	7973	7980	7987
63	·7993	8000	8007	8014	8021	8028	8035	8041	8048	8055
64	·8062	8069	8075	8082	8089	8096	8102	8109	8116	8122
65	·8129	8136	8142	8149	8156	8162	8169	8176	8182	8189
66	·8195	8202	8209	8215	8222	8228	8235	8241	8248	8254
67	·8261	8267	8274	8280	8287	8293	8299	8306	8312	8319
68	·8325	8331	8338	8344	8351	8357	8363	8370	8376	8382
69	·8388	8395	8401	8407	8414	8420	8426	8432	8439	8445
70	·8451	8457	8463	8470	8476	8482	8488	8494	8500	8506
71	·8513	8519	8525	8531	8537	8543	8549	8555	8561	8567
72	·8573	8579	8585	8591	8597	8603	8609	8615	8621	8627
73	·8633	8639	8645	8651	8657	8663	8669	8675	8681	8686
74	·8692	8698	8704	8710	8716	8722	8727	8733	8739	8745

	0	1	2	3	4	5	6	7	8	9
75	·8751	8756	8762	8768	8774	8779	8785	8791	8797	8802
76	·8808	8814	8820	8825	8831	8837	8842	8848	8854	8859
77	·8865	8871	8876	8882	8887	8893	8899	8904	8910	8915
78	·8921	8927	8932	8938	8943	8949	8954	8960	8965	8971
79	·8976	8982	8987	8993	8998	9004	9009	9015	9020	9025
80	·9031	9036	9042	9047	9053	9058	9063	9069	9074	9079
81	·9085	9090	9096	9101	9106	9112	9117	9122	9128	9133
82	·9138	9143	9149	9154	9159	9165	9170	9175	9180	9186
83	·9191	9196	9201	9206	9212	9217	9222	9227	9232	9238
84	·9243	9248	9253	9258	9263	9269	9274	9279	9284	9289
85	·9294	9299	9304	9309	9315	9320	9325	9330	9335	9340
86	·9345	9350	9355	9360	9365	9370	9375	9380	9385	9390
87	·9395	9400	9405	9410	9415	9420	9425	9430	9435	9440
88	·9445	9450	9455	9460	9465	9469	9474	9479	9484	9489
89	·9494	9499	9504	9509	9513	9518	9523	9528	9533	9538
90	·9542	9547	9552	9557	9562	9566	9571	9576	9581	9586
91	·9590	9595	9600	9605	9609	9614	9619	9624	9628	9633
92	·9638	9643	9647	9652	9657	9661	9666	9671	9675	9680
93	·9685	9689	9694	9699	9703	9708	9713	9717	9722	9727
94	·9731	9736	9741	9745	9750	9754	9759	9763	9768	9773
95	·9777	9782	9786	9791	9795	9800	9805	9809	9814	9818
96	·9823	9827	9832	9836	9841	9845	9850	9855	9855	9863
97	·9868	9872	9877	8881	9886	9890	9894	9899	9903	9908
98	·9912	9917	9921	9926	9930	9934	9939	9943	9948	9952
99	·9956	9961	9965	9969	9974	9978	9983	9987	9991	9996

ANTI-LOGARITHMS

	0	1	2	3	4	5	6	7	8	9
00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021
01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045
·02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069
·03	1072	1074	1076	1079	1081	1084	1086	1089	1091	1094
·04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119
·05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146
·06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172
·07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199
·08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227
·09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256
·10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285
·11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315
·12	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346
·13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377
·14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409
·15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442
·16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476
·17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510
·18	1514	1517	1521	1524	1528	1531	1535	1538	1542	1545
·19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581
·20	1585	1589	1592	1596	1600	1603	1607	1611	1614	1618
·21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656
·22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694
·23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734
·24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774
·25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816
·26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858
·27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901
·28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945
·29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991
·30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037
·31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084
·32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133
33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183

ANTI-LOGARITHMS

	0	1	2	3	4	5	6	7	8	9
·34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234
·35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286
·36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339
·37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393
·38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449
·39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506
·40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564
·41	2570	2576	2582	2588	2494	2600	2606	2612	2618	2624
·42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685
·43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748
·44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812
·45	2818	2825	2831	2838	2844	2851	2858	2864	2871	2877
·46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944
·47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013
·48	3020	3027	3034	3041	3048	3055	3062	3069	3076	3083
·49	3090	3097	3105	3112	3119	3126	3133	3141	3148	3155
·50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228
·51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304
·52	3311	3319	3327	3334	3342	3350	3357	3365	3373	3381
·53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459
·54	3467	3475	3483	3491	3499	3508	3516	3524	3532	3540
·55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622
·56	3631	3639	3648	3656	3664	3673	3681	3690	3698	3707
·57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793
·58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882
·59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972
·60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064
·61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159
·62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256
·63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355
·64	4365	4375	4385	4395	4406	4416	4426	4436	4446	4457
·65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560
·66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667

ANTI-LOGARITHMS

	0	1	2	3	4	5	6	7	8	9
·67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775
·68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887
·69	4898	4909	4920	4932	4943	4955	4966	4977	4989	5000
·70	5012	5023	5035	5047	5058	5070	5082	5093	5105	5117
·71	5129	5140	5152	5164	5176	5188	5200	5212	5224	5236
·72	5248	5260	5272	5284	5297	5309	5321	5333	5346	5358
·73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483
·74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610
·75	5623	5636	5649	5662	5675	5689	5702	5715	5728	5741
·76	5754	5768	5781	5791	5808	5821	5834	5848	5861	5875
·77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6012
·78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6152
·79	6166	6180	6194	6209	6223	6237	6252	6266	6281	6295
·80	6310	6324	6339	6353	6368	6383	6397	6412	6427	6442
·81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592
·82	6607	6622	6637	6653	6668	6683	6699	6714	6730	6745
·83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902
·84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063
·85	7079	7096	7112	7129	7145	7161	7178	7194	7211	7228
·86	7244	7261	7278	7295	7311	7328	7345	7362	7379	7396
·87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568
·88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745
·89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925
·90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110
·91	8128	8147	8166	8185	8204	8222	8241	8260	8279	8299
·92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492
·93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690
·94	8710	8730	8750	8770	8790	8810	8831	8851	8872	8892
·95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099
·96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311
·97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528
·98	9550	9572	9594	9616	9638	9661	9683	9705	9727	9750
·99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977

	0	1	2	3	4	5	6	7	8	9
10	1000	1020	1040	1061	1082	1103	1124	1145	1166	1188
11	1210	1232	1254	1277	1300	1323	1346	1369	1392	1416
12	1440	1464	1488	1513	1538	1563	1588	1613	1638	1664
13	1690	1716	1742	1769	1796	1823	1850	1877	1904	1932
14	1960	1988	2016	2045	2074	2103	2132	2161	2190	2220
15	2250	2280	2310	2341	2372	2403	2434	2465	2496	2528
16	2560	2592	2624	2657	2690	2723	2756	2789	2822	2856
17	2890	2924	2958	2993	3028	3063	3098	3133	3168	3204
18	3240	3276	3312	3349	3386	3423	3460	3497	3534	3572
19	3610	3648	3686	3725	3764	3803	3842	3881	3920	3960
20	4000	4040	4080	4121	4162	4203	4244	4285	4326	4368
21	4410	4452	4494	4537	4580	4623	4666	4709	4752	4796
22	4840	4884	4928	4973	5018	5063	5108	5153	5198	5244
23	5290	5336	5382	5429	5476	5523	5570	5617	5664	5712
24	5760	5808	5856	5905	5954	6003	6052	6101	6150	6200
25	6250	6300	6350	6401	6452	6503	6554	6605	6656	6708
26	6760	6812	6864	6917	6970	7023	7076	7129	7182	7236
27	7290	7344	7398	7453	7508	7563	7618	7673	7728	7784
28	7840	7896	7952	8009	8066	8123	8180	8237	8294	8352
29	8410	8468	8526	8585	8644	8703	8762	8821	8880	8940
30	9000	9060	9120	9181	9242	9303	9364	9425	9486	9548
31	9610	9672	9734	9797	9860	9923	9986			
31								1005	1011	1018
32	1024	1030	1037	1043	1050	1056	1063	1069	1076	1082
33	1089	1096	1102	1109	1116	1122	1129	1136	1142	1149
34	1156	1163	1170	1176	1183	1190	1197	1204	1211	1218
35	1225	1232	1239	1246	1253	1260	1267	1274	1282	1289
36	1296	1303	1310	1318	1325	1332	1340	1347	1354	1362
37	1369	1376	1384	1391	1399	1406	1414	1421	1429	1436
38	1444	1452	1459	1467	1475	1482	1490	1498	1505	1513
39	1521	1529	1537	1544	1552	1560	1568	1576	1584	1592
40	1600	1608	1616	1624	1632	1640	1648	1656	1665	1673

The position of the decimal point must be determined in accordance with the instructions given in the beginning.

SQUARES

	0	1	2	3	4	5	6	7	8	9
41	1681	1689	1697	1706	1714	1722	1731	1739	1747	1756
42	1764	1772	1781	1789	1798	1806	1815	1823	1832	1840
43	1849	1858	1866	1875	1884	1892	1901	1910	1918	1927
44	1936	1945	1954	1962	1971	1980	1989	1998	2007	2016
45	2025	2034	2043	2052	2061	2070	2079	2088	2098	2107
46	2116	2125	2134	2144	2153	2162	2172	2181	2190	2200
47	2209	2218	2228	2237	2247	2256	2266	2275	2285	2294
48	2304	2314	2323	2333	2343	2352	2362	2372	2381	2391
49	2401	2411	2421	2430	2440	2450	2460	2470	2480	2490
50	2500	2510	2520	2530	2540	2550	2560	2570	2581	2591
51	2601	2611	2621	2632	2642	2652	2663	2673	2683	2694
52	2704	2714	2725	2735	2746	2756	2767	2777	2788	2798
53	2809	2820	2830	2841	2852	2862	2873	2884	2894	2905
54	2916	2927	2938	2948	2959	2970	2981	2992	3003	3014
55	3025	3036	3047	3058	3069	3080	3091	3102	3114	3125
56	3136	3147	3158	3170	3181	3192	3204	3215	3226	3238
57	3249	3260	3272	3283	3295	3306	3318	3329	3341	3352
58	3364	3376	3387	3399	3411	3422	3434	3446	3457	3469
59	3481	3493	3505	3516	3528	3540	3552	3564	3576	3588
60	3600	3612	3624	3636	3648	3660	3672	3684	3697	3709
61	3721	3733	3745	3758	3770	3782	3795	3807	3819	3832
62	3844	3856	3869	3881	3894	3906	3919	3931	3944	3956
63	3969	3982	3994	4007	4020	4032	4045	4058	4070	4083
64	4096	4109	4122	4134	4447	4160	4173	4186	4199	4212
65	4225	4238	4251	4264	4277	4290	4303	4316	4330	4343
66	4356	4369	4382	4396	4409	4422	4436	4449	4462	4476
67	4489	4502	4516	4529	4543	4556	4570	4583	4597	4610
68	4624	4638	4651	4665	4679	4692	4706	4720	4733	4747
69	4761	4775	4789	4802	4816	4830	4844	4858	4872	4886
70	4900	4914	4928	4942	4956	4970	4984	4998	5013	5027
71	5041	5055	5069	5084	5098	5112	5127	5141	5155	5170
72	5184	5198	5213	5227	5242	5256	5271	5285	5300	5314
73	5329	5344	5358	5373	5388	5402	5417	5432	5446	5461

SQUARES

	0	1	2	3	4	5	6	7	8	9
74	5476	5491	5506	5520	5535	5550	5565	5580	5595	5610
75	5625	5640	5655	5670	5685	5700	5715	5730	5746	5761
76	5776	5791	5806	5822	5837	5852	5868	5883	5898	5914
77	5929	5944	5960	5975	5991	6006	6022	6037	6053	6068
78	6084	6100	6115	6131	6147	6162	6178	6194	6209	6225
79	6241	6257	6273	6288	6304	6320	6336	6352	6368	6384
80	6400	6416	6432	6448	6464	6480	6496	6512	6529	6545
81	6561	6577	6593	6610	6626	6642	6659	6675	6691	6708
82	6724	6740	6757	6773	6790	6806	6823	6839	6856	6872
83	6889	6906	6922	6939	6956	6972	6989	7006	7022	7039
84	7056	7073	7090	7106	7123	7140	7157	7174	7191	7208
85	7225	7242	7259	7276	7293	7310	7327	7344	7362	7379
86	7396	7413	7430	7448	7465	7482	7500	7517	7534	7552
87	7569	7586	7604	7621	7639	7656	7674	7691	7709	7726
88	7744	7762	7779	7797	7815	7832	7850	7868	7885	7903
89	7921	7939	7957	7974	7992	8010	8028	8046	8064	8082
90	8100	8118	8136	8154	8172	8190	8208	8226	8245	8263
91	8281	8299	8317	8336	8354	8372	8391	8409	8427	8446
92	8464	8482	8501	8519	8538	8556	8575	8593	8612	8630
93	8649	8668	8686	8705	8724	8742	8761	8780	8798	8817
94	8836	8855	8874	8892	8911	8930	8949	8968	8987	9006
95	9025	9044	9063	9082	9101	9120	9139	9158	9178	9197
96	9216	9235	9254	9274	9293	9312	9332	9351	9370	9390
97	9409	9428	9448	9467	9487	9506	9526	9545	9565	9584
98	9604	9624	9643	9663	9683	9702	9722	9742	9761	9781
99	9801	9821	9841	9860	9880	9900	9920	9940	9960	9980

SQUARE ROOTS

	0	1	2	3	4	5	6	7	8	9
10	1000 3162	1005 3178	1010 3194	1015 3209	1020 3225	1025 3240	1030 3256	1034 3271	1039 3286	1044 3302
11	1049 3317	1054 3332	1058 3347	1063 3362	1068 3376	1072 3391	1077 3406	1082 3421	1086 3435	1091 3450
12	1095 3464	1100 3479	1105 3493	1109 3507	1114 3521	1118 3536	1122 3550	1127 3564	1131 3578	1136 3592
13	1140 3606	1145 3619	1149 3633	1153 3647	1158 3661	1162 3674	1166 3688	1170 3701	1175 3715	1179 3728
14	1183 3742	1187 3755	1192 3768	1196 3782	1200 3795	1204 3808	1208 3821	1212 3834	1217 3847	1221 3860
15	1225 3873	1229 3886	1233 3899	1237 3912	1241 3924	1245 3937	1249 3950	1253 3962	1257 3975	1261 3987
16	1265 4000	1269 4012	1273 4025	1277 4037	1281 4050	1285 4062	1288 4074	1292 4087	1296 4099	1300 4111
17	1304 4123	1308 4135	1311 4147	1315 4159	1319 4171	1323 4183	1327 4195	1330 4207	1334 4219	1338 4231
18	1342 4243	1345 4254	1349 4266	1353 4278	1356 4290	1360 4301	1364 4313	1367 4324	1371 4336	1375 4347
19	1378 4359	1382 4370	1386 4382	1389 4393	1393 4405	1396 4416	1400 4427	1404 4438	1407 4450	1411 4461
20	1414 4472	1418 4483	1421 4494	1425 4506	1428 4517	1432 4528	1435 4539	1439 4550	1442 4561	1446 4572
21	1449 4583	1453 4593	1456 4604	1459 4615	1463 4626	1466 4637	1470 4648	1473 4658	1476 4669	1480 4680
22	1483 4690	1487 4701	1490 4712	1493 4722	1497 4733	1500 4743	1503 4754	1507 4764	1510 4775	1513 4785
23	1517 4796	1520 4806	1523 4817	1526 4827	1530 4837	1533 4848	1536 4858	1539 4868	1543 4879	1546 4889
24	1549 4899	1552 4909	1556 4919	1559 4930	1562 4940	1565 4950	1568 4960	1572 4970	1575 4980	1578 4990

The first significant figure and the position of the decimal point must be determined in accordance with the instructions given in the beginning

SQUARE ROOTS

	0	1	2	3	4	5	6	7	8	9
25	1581 5000	1584 5010	1587 5020	1591 5030	1594 5040	1597 5050	1600 5060	1603 5070	1606 5079	1609 5089
26	1612 5099	1616 5109	1619 5119	1622 5128	1625 5138	1628 5148	1631 5158	1634 5167	1637 5177	1640 5187
27	1643 5196	1646 5206	1649 5215	1652 5225	1655 5235	1658 5244	1661 5254	1664 5263	1667 5273	1670 5282
28	1673 5292	1676 5301	1679 5310	1682 5320	1685 5329	1688 5339	1691 5348	1694 5357	1697 5367	1700 5376
29	1703 5385	1706 5394	1709 5404	1712 5413	1715 5422	1718 5431	1720 5441	1723 5450	1726 5459	1729 5468
30	1732 5477	1735 5486	1738 5495	1741 5505	1744 5514	1746 5523	1749 5532	1752 5541	1755 5550	1758 5559
31	1761 5568	1764 5577	1766 5586	1769 5595	1772 5604	1775 5612	1778 5621	1780 5630	1783 5639	1786 5648
32	1789 5657	1792 5666	1794 5675	1797 5683	1800 5692	1803 5701	1806 5710	1808 5718	1811 5727	1814 5736
33	1817 5745	1819 5753	1822 5762	1825 5771	1828 5779	1830 5788	1833 5797	1836 5805	1838 5814	1841 5822
34	1844 5831	1847 5840	1849 5848	1852 5857	1855 5865	1857 5874	1860 5882	1863 5891	1865 5899	1868 5908
35	1871 5916	1873 5925	1876 5933	1879 5941	1881 5950	1884 5958	1887 5967	1889 5975	1892 5983	1895 5992
36	1897 6000	1900 6008	1903 6017	1905 6025	1908 6033	1910 6042	1913 6050	1916 6058	1918 6066	1921 6075
37	1924 6083	1926 6091	1929 6099	1931 6107	1934 6116	1936 6124	1939 6132	1942 6140	1944 6148	1947 6156
38	1949 6164	1952 6173	1954 6181	1957 6189	1960 6197	1962 6205	1965 6213	1967 6221	1970 6229	1972 6237
39	1975 6245	1977 6253	1980 6261	1982 6269	1985 6277	1987 6285	1990 6293	1992 6303	1995 6309	1997 6317
40	2000 6325	2002 6332	2005 6340	2007 6348	2010 6356	2012 6364	2015 6372	2017 6380	2020 6387	2022 6395

	0	1	2	3	4	5	6	7	8	9
41	2025 6403	2027 6411	2030 6419	2032 6427	2035 6434	2037 6442	2040 6450	2042 6458	2045 6465	2047 6473
42	2049 6481	2052 6488	2054 6496	2057 6504	2059 6512	2062 6519	2064 6527	2066 6535	2069 6542	2071 6550
43	2074 6557	2076 6565	2078 6573	2081 6580	2083 6588	2086 6595	2088 6603	2090 6611	2093 6618	2095 6626
44	2098 6633	2100 6641	2102 6648	2105 6656	2107 6663	2110 6671	2112 6678	2114 6686	2117 6693	2119 6701
45	2121 6708	2124 6716	2126 6723	2128 6731	2131 6738	2133 6745	2135 6753	2138 6760	2140 6768	2142 6775
46	2145 6782	2147 6790	2149 6797	2152 6804	2154 6812	2156 6819	2159 6826	2161 6834	2163 6841	2166 6848
47	2168 6856	2170 6863	2173 6870	2175 6877	2177 6885	2179 6892	2182 6899	2184 6907	2186 6914	2189 6921
48	2191 6928	2193 6935	2195 6943	2198 6950	2200 6957	2202 6964	2205 6971	2207 6979	2209 6986	2211 6993
49	2214 7000	2216 7007	2218 7014	2220 7021	2223 7029	2225 7036	2227 7043	2229 7050	2232 7057	2234 7064
50	2236 7071	2238 7078	2241 7085	2243 7092	2245 7099	2247 7106	2249 7113	2252 7120	2254 7127	2256 7134
51	2258 7141	2261 7148	2263 7155	2265 7162	2267 7169	2269 7176	2272 7183	2274 7190	2276 7197	2278 7204
52	2280 7211	2283 7218	2285 7225	2287 7232	2289 7239	2291 7246	2293 7253	2296 7259	2298 7266	2300 7273
53	2302 7280	2304 7287	2307 7294	2309 7301	2311 7308	2313 7314	2315 7321	2317 7328	2319 7335	2322 7342
54	2324 7348	2326 7355	2328 7362	2330 7369	2332 7376	2335 7382	2337 7389	2339 7396	2341 7403	2343 7409
55	2345 7416	2347 7423	2349 7430	2352 7436	2354 7443	2356 7450	2358 7457	2360 7468	2362 7470	2364 7477
56	2366 7483	2369 7490	2371 7497	2373 7503	2375 7510	2377 7517	2379 7523	2381 7530	2383 7537	2385 7543

	0	1	2	3	4	5	6	7	8	9
57	2387 7550	2390 7556	2392 7563	2394 7570	2396 7576	2398 7583	2400 7589	2402 7596	2404 7603	2406 7609
58	2408 7616	2410 7622	2412 7629	2415 7635	2417 7642	2419 7649	2421 7655	2423 7662	2425 7668	2427 7675
59	2429 7681	2431 7688	2433 7694	2435 7701	2437 7707	2439 7714	2441 7720	2443 7727	2445 7733	2447 7740
60	2449 7746	2452 7752	2454 7759	2456 7765	2458 7772	2460 7778	2462 7785	2464 7791	2466 7797	2468 7804
61	2470 7810	2472 7817	2474 7823	2476 7829	2478 7836	2480 7842	2482 7849	2484 7855	2486 7861	2488 7868
62	2490 7874	2492 7880	2494 7887	2496 7893	2498 7899	2500 7906	2502 7912	2504 7918	2506 7925	2508 7931
63	2510 7937	2512 7944	2514 7950	2516 7956	2518 7962	2520 7969	2522 7975	2524 7981	2526 7987	2528 7994
64	2530 8000	2532 8006	2534 8012	2536 8019	2538 8025	2540 8031	2542 8037	2544 8044	2546 8050	2548 8056
65	2550 8062	2551 8068	2553 8075	2555 8081	2557 8087	2559 8093	2561 8099	2563 8106	2565 8112	2567 8118
66	2569 8124	2571 8130	2573 8136	2575 8142	2577 8149	2579 8155	2581 8161	2583 8107	2585 8173	2587 8179
67	2588 8185	2590 8191	2592 8198	2594 8204	2596 8210	2598 8216	2600 8222	2602 8228	2604 8234	2606 8240
68	2608 8246	2610 8252	2612 8258	2613 8264	2615 8270	2617 8276	2619 8283	2621 8289	2623 8295	2625 8301
69	2627 8307	2629 8313	2631 8319	2632 8325	2634 8331	2636 8337	2638 8843	8640 8349	2642 8355	2644 8361
70	2646 8367	2648 8373	2650 8379	2651 8385	2653 8390	2655 8396	2657 8402	2659 8408	2661 8414	2663 8420
71	2665 8426	2666 8432	2668 8438	2670 8444	2672 8450	2674 8456	2676 8462	2678 8468	2680 8473	2681 8479
72	2683 8485	2685 8491	2687 8497	2689 8503	2691 8509	2693 8515	2694 8521	2696 8526	2698 8532	2700 8538

SQUARE ROOTS

	0	1	2	3	4	5	6	7	8	9
73	2702 8544	2704 8550	2706 8556	2707 8562	2709 8567	2711 8573	2713 8579	2715 8585	2717 8591	2718 8597
74	2720 8602	2722 8608	2724 8614	2726 8620	2728 8626	2729 8631	2731 8637	2733 8643	2735 8649	2737 8654
75	2739 8660	2740 8666	2742 8672	2744 8678	2746 8683	2748 8689	2750 8695	2751 8701	2753 8706	2755 8712
76	2757 8718	2759 8724	2760 8729	2762 8735	2764 8741	2766 8746	2768 8752	2769 8758	2771 8764	2773 8769
77	2775 8775	2777 8781	2778 8786	2780 8792	2782 8798	2784 8803	2786 8809	2787 8815	2789 8820	2791 8826
78	2793 8832	2795 8837	2796 8843	2798 8849	2800 8854	2802 8860	2804 8866	2805 8871	2807 8877	2809 8883
79	2811 8888	2812 8894	2814 8899	2816 8905	2818 8911	2820 8916	2821 8922	2823 8927	2825 8933	2827 8939
80	2828 8944	2830 8950	2832 8955	2834 8961	2835 8967	2837 8972	2839 8978	2841 8983	2843 8989	2844 8994
81	2846 9000	2848 9006	2850 9011	2851 9017	2853 9022	2855 9028	2857 9033	2858 9039	2860 9044	2862 9050
82	2864 9055	2865 9061	2867 9066	2869 9072	2871 9077	2872 9083	2874 9088	2876 9094	2877 9099	2879 9105
83	2881 9110	2883 9116	2884 9121	2886 9127	2888 9132	2890 9138	2891 9143	2893 9149	2895 9154	2897 9160
84	2898 9165	2900 9171	2902 9176	2903 9182	2905 9187	2907 9192	2909 9198	2910 9203	2912 9209	2914 9214
85	2915 9220	2917 9225	2919 9230	2921 9236	2922 9241	2924 9247	2926 9252	2927 9257	2929 9263	2931 9268
86	2933 9274	2934 9279	2936 9284	2938 9290	2939 9295	2941 9301	2943 9306	2944 9311	2946 9317	2948 9322
87	2950 9327	2951 9333	2953 9338	2955 9343	2956 9349	2958 9354	2960 9359	2961 9365	2963 9370	2965 9375
88	2966 9381	2968 9386	2970 9391	2972 9397	2973 9402	2975 9407	2977 9413	2978 9418	2980 9423	2982 9429

	0	1	2	3	4	5	6	7	8	9
89	2983 9434	2985 9439	2987 9445	2988 9450	2990 9455	2992 9460	2993 9466	2995 9471	2997 9476	2998 9482
90	3000 9487	3002 9492	3003 9497	3005 9503	3007 9508	3008 9513	3010 9518	3012 9524	3013 9529	3015 9534
91	3017 9539	3018 9545	3020 9550	3022 9555	3023 9560	3025 9566	3027 9571	3028 9576	3030 9581	3032 9586
92	3033 9592	3035 9597	3036 9602	3038 9607	3040 9612	3041 9618	3043 9623	3045 9628	3046 9633	3048 9638
93	3050 9644	3051 9649	3053 9654	3055 9659	3056 9664	3058 9670	3059 9675	3061 9680	3063 9685	3064 9690
94	3066 9695	3068 9701	3069 9706	3071 9711	3072 9716	3074 9721	3076 9726	3077 9731	3079 9737	3081 9742
95	3082 9747	3084 9752	3085 9757	3087 9662	3089 9767	3090 9772	3092 9778	3094 9783	3095 9788	3097 9793
96	3098 9798	3100 9803	3102 9808	3103 9813	3105 9818	3106 9823	3108 9829	3110 9834	3111 9839	3113 9844
97	3114 9849	3116 9854	3118 9859	3119 9864	3121 9869	3122 9874	3124 9879	3126 9884	3127 9889	3129 9894
98	3130 9899	3132 9905	3134 9910	3135 9915	3137 9920	3138 9925	3140 9930	3142 9935	3143 9940	3145 9945
99	3146 9950	3148 9955	3150 9960	3151 9965	3153 9970	3154 9975	3156 9980	3158 9985	3159 9990	3161 9995

RECIPROCALS

	0	1	2	3	4	5	6	7	8	9
1·0	1·0000	·9901	·9804	·9709	·9615	·9524	·9434	·9346	·9259	·9174
1·1	·9091	·9009	·8929	·8850	·8772	·8696	·8621	·8547	·8475	·8403
1·2	·8333	·8264	·8197	·8130	·8065	·8000	·7937	·7874	·7813	·7752
1·3	·7692	·7634	·7576	·7519	·7463	·7407	·7353	·7299	·7246	·7194
1·4	·7143	·7092	·7042	·6993	·6944	·6897	·6849	·6803	·6757	·6711
1·5	·6667	·6623	·6579	·6536	·6494	·6452	·6410	·6369	·6329	·6289
1·6	·6250	·6211	·6173	·6135	·6098	·6061	·6024	·5988	·5952	·5917
1·7	·5882	·5848	·5814	·5780	·5747	·5714	·5682	·5650	·5618	·5587
1·8	·5556	·5525	·5495	·5464	·5435	·5405	·5376	·5348	·5319	·5291
1·9	·5263	·5236	·5208	·5181	·5155	·5128	·5102	·5076	·5051	·5025
2·0	·5000	·4975	·4950	·4926	·4902	·4878	·4854	·4831	·4808	·4785
2·1	4762	·4739	·4717	·4695	·4673	·4651	·4630	·4608	·4587	·4566
2·2	·4545	·4525	·4505	·4484	·4464	·4444	·4425	·4405	·4386	·4367
2·3	·4348	·4329	·4310	·4292	·4274	·4255	·4237	·4219	·4202	·4184
2·4	·4167	·4149	·4132	·4115	·4098	·4082	·4065	·4049	·4032	·4016
2·5	·4000	·3984	·3968	·3953	·3937	·3922	·3906	·3891	3876	·3861
2·6	·3846	·3831	·3817	·3802	·3788	·3774	·3759	·3745	·3731	·3717
2·7	·3704	·3690	·3676	·3663	·3650	·3636	·3623	·3610	·3597	·3584
2·8	·3571	·3559	·3546	·3534	·3521	·3509	·2497	·3484	·3472	·3460
2·9	·3448	·3436	·3425	·3413	·3401	·3390	·3378	·3367	·3356	·3344
3·0	·3333	·3322	·3311	·3300	·3289	·3279	·3268	·3257	·3247	·3236
3·1	·3226	·3215	·3205	·3195	·3185	·3175	·3165	·3155	·3145	·3135
3·2	·3125	·3115	·3106	·3096	·3086	·3077	·3067	·3058	·3049	·3040
3·3	·3030	·3021	·3012	·3003	·2994	·2985	·2976	·2967	·2959	·2950
3·4	·2941	·2933	·2924	·2915	·2907	·2899	·2890	·2882	·2874	·2865
3·5	·2857	·2849	·2841	·2833	·2825	·2817	·2809	·2801	·2793	·2786
3·6	·2778	·2770	·2762	·2755	·2747	·2740	·2732	·2725	·2717	·2710
3·7	·2703	·2695	·2688	·2681	·2674	·2667	·2660	·2653	·2646	·2639
3·8	·2632	·2625	·2618	·2611	·2604	·2597	·2591	·2584	·2577	·2571
3·9	·2564	·2558	·2551	·2545	·2538	·2532	·2525	·2519	·2513	·2506
4·0	·2500	·2494	·2488	·2481	·2475	·2469	·2463	·2457	·2451	·2445
4·1	·2439	·2433	·2427	·2421	·2415	·2410	·2404	2398	·2392	·2387

	0	1	2	3	4	5	6	7	8	9
4·2	2381	·2375	·2370	·2364	·2358	·2353	·2347	·2342	·2336	·2331
4·3	·2326	·2320	·2315	·2309	·2304	·2299	·2294	·2288	·2283	·2278
4·4	2273	·2268	·2262	·2257	·2252	·2247	·2242	·2237	·2232	·2227
4·5	·2222	·2217	·2212	·2208	·2203	·2198	·2193	·2188	·2183	·2179
4·6	·2174	·2169	·2165	·2160	·2155	·2151	·2146	·2141	·2137	·2132
4·7	·2128	·2123	·2119	·2114	·2110	·2105	·2101	·2096	·2092	·2088
4·8	·2083	·2079	·2075	·2070	·2066	·2062	·2058	·2053	·2049	·2045
4·9	·2041	·2037	·2033	·2028	2024	·2020	·2016	·2012	·2008	·2004
5·0	·2000	·1996	·1992	·1988	·1984	·1980	·1976	·1972	·1969	·1965
5·1	·1961	·1957	·1953	·1949	·1946	·1942	·1938	·1934	·1931	·1927
5·2	·1923	·1919	·1916	·1912	·1908	·1905	·1901	·1898	·1894	·1890
5·3	·1887	·1883	·1880	·1876	·1873	·1869	·1866	·1862	·1859	·1855
5·4	·1852	·1848	·1845	·1842	·1838	·1835	·1832	·1828	·1825	·1821
5·5	·1818	·1815	·1812	1808	·1805	·1802	·1799	·1795	·1792	·1789
5·6	·1786	·1783	·1779	·1776	·1773	·1770	·1767	·1764	·1761	·1757
5·7	·1754	·1751	·1748	·1745	·1742	·1739	·1736	·1733	·1730	·1727
5·8	·1724	·1721	·1718	·1715	·1712	·1709	·1706	·1704	·1701	·1698
5·9	·1695	·1692	·1689	·1686	·1684	·1681	·1678	·1675	·1672	·1669
6·0	·1667	·1664	·1661	·1658	·1656	·1653	·1650	·1647	·1645	·1642
6·1	·1639	·1637	·1634	·1631	·1629	·1626	·1623	·1621	·1618	·1616
6·2	·1613	·1610	·1608	·1605	·1603	·1600	·1597	·1595	·1592	·1590
6·3	·1587	·1585	·1582	·1580	·1577	·1575	·1572	·1570	·1567	·1565
6·4	·1563	·1560	·1558	·1555	·1553	·1550	·1548	·1546	·1543	·1541
6·5	·1538	·1536	·1534	·1531	·1529	·1527	·1524	·1522	·1520	·1517
6·6	·1515	·1513	·1511	·1508	·1506	·1504	·1502	·1499	·1497	·1495
6·7	·1493	·1490	·1488	·1486	·1484	·1481	·1479	·1477	·1475	·1473
6·8	·1471	·1468	·1466	·1464	·1462	·1460	·1458	·1456	·1453	·1451
6·9	·1449	·1447	·1445	·1443	·1441	·1439	·1437	·1435	·1433	·1431
7·0	·1429	·1427	·1425	·1422	·1420	·1418	·1416	·1414	·1412	·1410
7·1	·1408	·1406	·1404	·1403	·1401	·1399	·1397	·1395	·1393	·1391
7·2	·1389	·1387	·1385	·1383	·1381	·1379	·1377	1376	·1374	·1372
7·3	·1370	·1368	·1366	·1364	·1362	·1361	·1359	·1357	·1355	·1353
7·4	1351	·1350	·1348	·1346	·1344	·1342	·1340	·1339	·1337	·1335

RECIPROCALS

	0	1	2	3	4	5	6	7	8	9
7·5	·1333	·1332	·1330	·1328	·1326	·1325	·1323	·1321	·1319	·1318
7·6	·1316	·1314	·1312	·1311	·1309	·1307	·1305	·1304	·1302	·1300
7·7	·1299	·1297	·1295	·1294	·1292	·1290	·1289	·1287	·1285	·1284
7·8	·1282	·1280	·1279	·1277	·1276	·1274	·1272	·1271	·1269	·1267
7·9	·1266	·1264	·1263	·1261	·1259	·1258	1256	·1255	·1253	·1252
8·0	·1250	·1248	·1247	·1245	·1244	·1242	·1241	·1239	·1238	·1236
8·1	·1235	·1233	·1232	·1230	·1229	·1227	·1225	·1224	·1222	·1221
8·2	·1220	·1218	·1217	·1215	·1214	·1212	·1211	·1209	·1208	·1206
8·3	·1205	·1203	·1202	·1200	·1199	·1198	·1196	·1195	·1193	·1192
8·4	·1190	·1189	·1188	·1186	·1185	·1183	1182	·1181	·1179	·1178
8·5	·1176	·1175	·1174	·1172	·1171	·1170	·1168	·1167	·1166	·1164
8·6	·1163	·1161	·1160	·1159	·1157	·1156	·1155	·1153	·1152	·1151
8·7	·1149	·1148	·1147	·1145	·1144	·1143	·1142	·1140	·1139	·1138
8·8	·1136	·1135	·1134	·1133	·1131	·1130	·1129	·1127	·1126	·1125
8·9	·1124	·1122	·1121	·1120	·1119	·1117	·1116	·1115	·1114	·1112
9·0	·1111	·1110	·1109	·1107	·1106	·1105	·1104	·1103	·1101	·1100
9·1	·1099	·1098	·1096	·1095	·1094	·1093	1092	·1091	·1089	·1088
9·2	·1087	·1086	·1085	·1083	·1082	·1081	1080	·1079	·1078	·1076
9·3	·1075	·1074	·1073	·1072	·1071	·1070	1068	·1067	·1066	·1165
9·4	·1064	·1063	·1062	·1060	·1059	·1058	·1057	·1056	·1055	·1054
9·5	·1053	·1052	·1050	·1049	·1048	·1047	·1046	·1045	·1044	·1043
9·6	·1042	·1041	·1040	·1038	·1037	·1036	·1035	·1034	·1033	·1032
9·7	·1031	·1030	·1029	·1028	·1027	·1026	·1025	·1024	·1022	·1021
9·8	·1020	·1019	·1018	·1017	·1016	·1015	·1014	·1013	·1012	·1011
9·9	·1010	1009	·1008	·1007	·1006	·1005	·1004	·1003	·1002	·1001

EXTRACTS FROM SOME REVIEWS OF STATSTICS: THEORY AND PRACTICE

The book under review represents a successful attempt at synthesising the theoretical and practical aspects of the statistical method as applied to social and economic problems...... Two of the best chapters of the book are entitled *Statistical material in India* and *Indian and Foreign Index numbers*...... The most useful chapter of the book is the last which warns the student of the dangers attending on an indiscriminate use of statistical measures....the authors have succeeded in producing a "Mills" for Indian students and have gone one better by including a large number of exercises at the end of each chapter.

The Eastern Economist
February 4, 1944.

Messrs. Ghosh and Chaudhri have done a valuable service by drawing pointed attention in their book to the inadequate quality and content of official statistics in India......The chapters relating to Index numbers....are well written....An analysis of the index numbers prepared in England and America is also given and places the Indian effort in true perspective....The usefulness of the book is enhanced by appendices relating to the statistical publications in India, the Bowley Robertson Committee's scheme for measurement of the national income and mathematical tables.

The Commerce,
March 11, 1944.

The authors have throughout the book maintained a highly scientific attitude....The chapters on index numbers and correlation show much originality and freshness. Out of the existing text books on statistics for Indian students the book under review is by far the best and we heartily recommend it to university students and those members of the public who are interested in learning this subject.

The Indian Journal of Economics,
January, 1944.

In this handy volume the authors have successfully compressed the main principles of modern statistical science, both in its theoretical and applied aspect. On the whole it will serve as a useful introduction to the subject, while the simple and non-mathematical presentation is likely to render it attractive also to the layman.

The Capital
April 27, 1944.